GEOMETRIC COVERAGE EXPLAINS RECOVERY OF REINFORCEMENT LEARNING POLICIES FROM SYNTHETIC OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a trained reinforcement learning policy be recovered without trajectories, rewards, or valid environment states? Across control tasks, students trained only on teacher-labeled synthetic inputs recover substantial closed-loop performance. Recovery is not explained by physical realism alone. Instead, it depends on two geometric properties of the query distribution: excitation of teacher-sensitive directions and sufficient variance outside the dominant sensitive subspace. Matched subspace interventions isolate directional excitation, while orthogonal thickening isolates ambient support; on Humanoid-v5, 0.10-scale orthogonal noise raises recovery from 8.1% to 73.7% with almost unchanged sensitivity overlap. We show theoretically that covariance mismatch governs transfer from query-space error to deployment error and that strict subspace queries can leave deployment behavior unidentifiable. A synthetic-only pilot further shows that inverse-covariance coverage can diagnose failures missed by raw overlap, although it is not universally predictive. These effects replicate for SAC and discrete-action ReLU policies. Small real-state buffers provide complementary deployment information, and action-imitation accuracy can continue improving after closed-loop return saturates. Together, these results identify geometric coverage as a central factor governing passive recovery of state-based policies.

1 INTRODUCTION

A trained reinforcement learning (RL) actor is ultimately a function from observations to actions. We ask a deliberately narrow structural question: how much of that function can be recovered when a student sees no trajectories, rewards, transitions, or valid environment states?

We study a stricter setting than standard imitation, policy distillation, or "data-free" knowledge transfer. The student receives independent inputs from a fixed elementary distribution and the corresponding teacher actions. The query distribution is chosen before observing teacher outputs and is not adapted using teacher activations, gradients, confidence, student disagreement, rollouts, rewards, or environment data. Thus, the student is a probe of the trained policy function rather than a replacement for interactive imitation. Section 2 positions this restriction against data-free knowledge distillation, model extraction, and policy distillation.

The result is surprising. Across state-based control tasks, teacher-labeled Gaussian, Laplace, and uniform inputs can recover substantial closed-loop performance even when the inputs are physically invalid. Physical validity is therefore not necessary and, by itself, does not explain recovery: some broad invalid distributions succeed while other distributions with comparable coordinate scale fail.

We test a geometric explanation through controlled interventions. The teacher sensitivity matrix $G_{\pi_T} = \mathbb{E}[J_{\pi_T}(x)^\top J_{\pi_T}(x)]$ identifies directions along which the action map changes. For a synthetic distribution with covariance Σ_q, the raw sensitivity-overlap score $\mathcal{C}(q) = \mathrm{Tr}(G_{\pi_T}\Sigma_q)$ measures excitation of those directions. At matched support dimension and per-direction variance, top-sensitive subspaces outperform random and bottom-sensitive subspaces. Overlap alone is not sufficient, however. A strict top subspace leaves the student unconstrained in orthogonal directions encountered during rollout. Adding only 0.10-scale orthogonal noise raises Humanoid-v5 recovery from 8.1% to 73.7% while leaving raw overlap almost unchanged. At the other extreme, a full isotropic Gaussian

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.

has maximal raw overlap but spends much more variance outside the sensitive subspace and recovers less well than lightly thickened queries. The intervention therefore separates two requirements: sensitivity-directed excitation and sufficient ambient support.

We formalize why insufficient ambient support can prevent query-space agreement from transferring to deployment. Under a local linear error model, rollout error is bounded by the largest covariance mismatch between deployment and query distributions, while strict subspace queries admit no finite general bound. We also test whether geometry can be found without rollouts. A 1,000-query Gaussian pilot estimates a teacher-sensitive subspace using only synthetic Jacobians. The resulting thickened distribution does not reliably outperform isotropic sampling. Its failures are predicted by $\mathcal{R}_{\text{inv}} = \text{Tr}(\Sigma_\rho \Sigma_q^{-1})$ on post-design rollout diagnostics, while raw subspace overlap is not. Thus, isotropic sampling remains a strong minimax baseline when deployment geometry is unknown.

Small real buffers provide a complementary signal. Under a fixed label budget, 5k–20k rollout observations sharply improve harder environments, while contiguous and independently sampled rollout states perform similarly. This shows that the marginal state distribution, rather than temporal order, is the main benefit in these experiments. We interpret real states as rollout-region anchors and synthetic queries as broader functional constraints. The synthetic distribution itself is not re-centered on those anchors, so we do not claim that the experiment causally demonstrates localized synthetic sampling.

The same directional ordering and thickening effect replicate in a discrete ReLU PPO policy and an off-policy SAC actor. A controlled MNIST study provides a boundary: label memorization raises effective sensitivity rank while synthetic recoverability falls, despite nearly unchanged training accuracy. These checks indicate that low-dimensional sensitivity can help recovery but is not sufficient when the learned function becomes locally irregular or memorization-dominated. Our claims remain restricted to compact vector-input networks.

Our contributions are:

- **A controlled recovery phenomenon.** Passive, teacher-labeled synthetic inputs recover substantial closed-loop policy performance without trajectories, rewards, or environment interaction.
- **Controlled evidence for a geometric mechanism.** Matched subspace interventions show that teacher-sensitive directions carry more useful supervision, while orthogonal thickening isolates the separate need for ambient support.
- **Coverage guarantees and a synthetic-only test.** We prove strict-subspace non-identifiability and covariance-transfer bounds, then test synthetic-only geometry design across four environments. Inverse-covariance coverage diagnoses the principal pilot failures, but neither it nor raw sensitivity overlap is a universal predictor of recovery.
- **A real-anchor result.** Fixed-budget mixing and trajectory-order controls show that small real buffers can strongly improve recovery and that their benefit is primarily associated with the visited-state marginal, not temporal order.
- **Breadth and boundaries.** Off-policy, discrete-action, supervised-learning, perturbation, and constrained-control studies show where the mechanism transfers. We also show that action imitation can continue to improve after closed-loop return has saturated, separating function recovery from control recovery.

2 Related Work

Imitation and policy distillation. Behavioral cloning, DAgger (Ross et al., 2011), adversarial imitation, and policy distillation train on states obtained from demonstrations, expert or student rollouts, or direct environment interaction (Pomerleau, 1991; Ross et al., 2011; Ho & Ermon, 2016; Rusu et al., 2015; Czarnecki et al., 2019). Offline variants retain valid demonstrations or transition datasets (Fu et al., 2020; Ma et al., 2022; Kim et al., 2021; Xu et al., 2022; Ren et al., 2024). Behaviour distillation can compress state-action data through an outer objective that includes policy training and environment evaluation (Lupu et al., 2024). In contrast, our fully synthetic experiments do not sample from a rollout distribution and do not use environment feedback.

Data-free knowledge transfer. Data-free and zero-shot knowledge distillation remove access to the original training set, but typically reconstruct informative surrogate inputs from the teacher. Prior methods use stored activation statistics or metadata (Lopes et al., 2017), optimize inputs to match teacher-derived class structure (Nayak et al., 2019), invert internal feature statistics (Yin et al., 2019), or train generators to elicit strong or adversarial teacher responses (Chen et al., 2019; Fang et al., 2019). These methods are zero-original-data, but their queries are deliberately engineered using the teacher or student. Our setting removes this reconstruction step: the query distribution is fixed independently of teacher outputs, activations, gradients, student disagreement, and environment data. The question is therefore not how to synthesize teacher-like data, but when unadapted synthetic probing is already sufficient to recover a policy function. Our separate synthetic-pilot study uses white-box Jacobians to test whether the paper's geometry can guide queries. It is not part of the passive baseline and does not use rollout data.

Model extraction and unrelated-data transfer. Black-box model extraction likewise trains a substitute from queried input-output pairs. Knockoff Nets showed that out-of-distribution natural images can recover useful classifier functionality, while later data-free extraction methods generate or actively select queries to improve fidelity and query efficiency (Orekondy et al., 2018; Truong et al., 2021). Our experiments study closed-loop control and intervene on the geometry of a fixed synthetic distribution rather than optimize an attack strategy. Recent subliminal-learning results show that teacher traits can transfer through semantically unrelated generated data (Cloud et al., 2025). That setting studies latent trait transmission, often between models sharing a base initialization; here every target is the teacher's direct action on the queried observation, and the goal is to explain functional recovery through input-space sensitivity and support.

Simulator-free transfer and sensitivity geometry. Prior work has established several neighboring forms of data-free or synthetic distillation, but under different query mechanisms. The Generative Adversarial Simulator (GAS) performs simulator-free RL policy distillation using a learned, adaptive generator to construct informative teacher queries (Raiman, 2020). Augmented Policy Cloning (APC) generates synthetic states around sampled expert trajectories, and therefore retains direct information about the expert's visited-state distribution (Galashov et al., 2022). Outside RL, data-free knowledge distillation has shown that arbitrary transfer sets, including random noise and semantically unrelated inputs, can support teacher–student transfer (Nayak et al., 2021), and Gaussian-noise distillation has been studied directly together with its failure modes (Raikwar & Mishra, 2022). Active-subspace methods have also been used to identify low-dimensional directions governing neural-network sensitivity (Cui et al., 2020).

In this paper, queries are sampled from a fixed exogenous distribution chosen independently of the teacher, student, and environment. There is no learned generator, adaptive query selection, rollout state, trajectory, reward, or environment interaction used to construct the synthetic inputs. To our knowledge, this fixed, non-adaptive querying regime has not previously been studied for closed-loop RL policy recovery. Our contribution is a characterization of when such uninformed queries are sufficient, building on prior work in synthetic distillation and sensitivity subspaces. Matched subspace and thickening interventions separate excitation of teacher-sensitive directions from ambient support, while covariance-based coverage results characterize when query-space agreement can transfer to the deployment distribution and when it cannot.

3 Synthetic Probing Setup

We study a trained reinforcement learning actor as a function from normalized policy inputs to actions. Let $\pi_T : \mathcal{X} \to \mathcal{A}$ denote a fixed teacher policy, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the normalized input space and $\mathcal{A} \subseteq \mathbb{R}^k$ is the action space. For stochastic actors, we use the deterministic action mean. The teacher is trained with standard reinforcement learning and remains fixed throughout all probing experiments. We train a student $\pi_S : \mathcal{X} \to \mathcal{A}$ to imitate the teacher using only synthetic inputs. Each MuJoCo experiment uses 10 independently trained PPO teachers. For every condition, each teacher is paired with one independently initialized student. Teachers and students are evaluated for 100 episodes using paired evaluation seeds.

During PPO training, observation normalization statistics $(\mu_{\mathrm{rms}}, \sigma_{\mathrm{rms}})$ are accumulated using the standard `VecNormalize` wrapper. These statistics are used only when executing the trained policies

in the environment: raw observations are normalized before being passed to the actor. They are not used to construct the synthetic training distributions.

All synthetic inputs are defined directly in the normalized policy-input space. Our baseline is $x \sim \mathcal{N}(0, I_d)$, and we also consider Laplace, uniform, narrow, wide, and shifted variants. These distributions are fixed independently of rollout data. They are not fitted to the teacher's state distribution, do not use empirical state covariances, and need not correspond to physically valid environment states.

Importantly, normalized rollout observations are not generally distributed as $\mathcal{N}(0, I_d)$. The normalization statistics are running estimates collected during training, while the state distribution changes as the policy improves. Real rollout inputs therefore remain correlated, only approximately centered, and generally non-Gaussian. The baseline $\mathcal{N}(0, I_d)$ distribution should thus be viewed as a simple probe of the actor's normalized input space, rather than an approximation to its rollout distribution.

Given a synthetic dataset $\mathcal{D}_{\text{syn}} = \{(x_i, a_i)\}_{i=1}^N, a_i = \pi_T(x_i)$, the student minimizes $\mathcal{L}(\pi_S) = \frac{1}{N}\sum_{i=1}^N \|\pi_S(x_i) - \pi_T(x_i)\|_2^2$. Student training requires no environment interaction. The teacher is queried only on independently sampled synthetic inputs, which are not optimized to match rollout states or internal teacher statistics. The pilot-guided experiment in Section 5.5 is the only exception to fixed passive sampling. There, we use teacher Jacobians computed at 1,000 Gaussian pilot inputs to construct a second synthetic query distribution. This experiment still uses no rollout data for either design or training, but it requires white-box access to the teacher and is therefore reported separately from the passive baseline.

We evaluate each trained student by executing it in the original environment under nominal dynamics. Raw observations are normalized using the stored `VecNormalize` statistics before being passed to the student, exactly as for the teacher. Let $R(\pi)$ denote the mean episodic return over 100 evaluation episodes. We report recovery relative to the teacher as $\text{Recovery}(\pi_S) = 100 \times \frac{R(\pi_S)}{R(\pi_T)}$. Recovery can exceed 100% because the teacher need not be the unique policy achieving its level of return under deterministic evaluation. All matched query-distribution comparisons use the same student architecture and optimization hyperparameters; we do not tune training separately for each distribution. This keeps optimization capacity fixed when isolating query geometry, although it does not imply that every distribution is individually optimized.

This setup deliberately removes the main forms of structure normally available in imitation learning or policy distillation: valid trajectories, temporal correlations, and training inputs chosen to resemble the environment state distribution. Successful recovery must therefore come from structure in the teacher policy as a function over normalized inputs, rather than from the physical validity of the synthetic states. Most data-free distillation methods therefore assume a richer teacher-access model than our passive setting, using model weights, internal statistics, gradients, or adaptive teacher/student feedback to construct queries. Our passive experiments require only input–output access to the fixed teacher; white-box access is introduced separately only in the synthetic-pilot study.

4 Recovery from Synthetic Observations

We first ask whether synthetic observations alone contain enough information to recover a trained policy.

Synthetic observations recover substantial teacher performance. Figure 1 shows that synthetic supervision recovers a large fraction of teacher performance across four MuJoCo locomotion tasks. Hopper-v5 recovers near-teacher performance using 10^5 synthetic samples, HalfCheetah-v5 reaches teacher-level recovery by 2.5×10^5 samples, and Ant-v5 approaches teacher performance at 2×10^6 samples. Humanoid-v5 is harder but still recovers a substantial fraction of teacher performance despite its 348-dimensional observation space. Full numerical results are reported in Appendix D.2.

The positive intuition is geometric rather than semantic. Synthetic inputs need not be valid states if they excite directions along which the teacher's action function varies and provide enough ambient support for closed-loop deployment. Section 5 formalizes this through sensitivity overlap and subspace interventions.

Recovery is not explained by physical validity alone. A natural explanation is that synthetic samples work only because they approximate the empirical state distribution. Figure 1 challenges this view: centered Gaussian and Laplace distributions perform well, and uniform sampling becomes competitive in Ant-v5 and Humanoid-v5 despite producing physically invalid inputs. Thus, semantic or physical validity is not required for useful supervision. The shifted distribution provides a complementary stress test. It can fail badly even when its coordinate-wise scale is comparable to successful centered distributions, showing that arbitrary synthetic distributions do not work. However, it is not a clean geometric control: in high-dimensional normalized space, shifting every coordinate by a constant greatly increases aggregate displacement and can push tanh networks toward saturation. We therefore do not use shifted-distribution failure alone as evidence against Euclidean proximity.

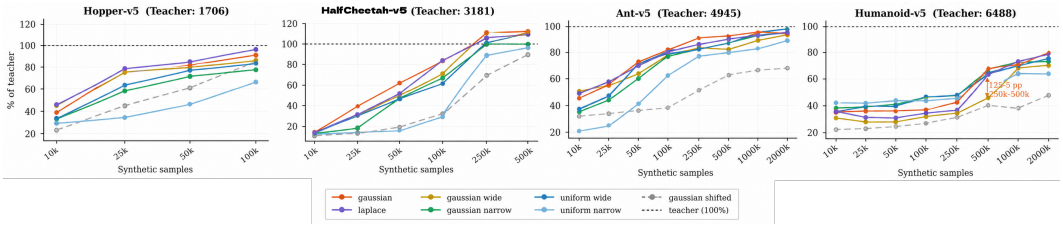

Figure 1: Synthetic recovery across MuJoCo environments. Recovery improves with sample count. The shifted Gaussian degrades performance relative to centered baselines, while uniform sampling becomes competitive in higher-dimensional tasks.

The stronger evidence comes from controls that separate Euclidean proximity from policy-relevant geometry. Uniform-wide sampling succeeds in high dimensions despite producing physically invalid observations far from the rollout distribution. Appendix D.12 shows that such samples can nevertheless retain substantial projections onto the teacher's dominant sensitivity directions, providing a concentration-of-measure explanation for why ambient Euclidean distance alone does not determine recovery. The real-fit Gaussian diagnostic in Appendix D.9 further shows that, in Humanoid-v5, matching covariance structure matters substantially more than recentring the query distribution. Most directly, the subspace intervention in Section 5.2 holds support dimension and per-direction variance fixed while changing only the sampled eigenspace; teacher-sensitive directions recover substantially more behavior than random or insensitive subspaces. Table 12 further addresses the shifted-baseline confound. A norm-controlled random shift with $c = 0.75$ still underperforms the centered Gaussian baseline, showing that the shifted-distribution gap is not solely an artifact of increased input norm or activation saturation. Scaling also reveals a coverage threshold: Hopper-v5 saturates at moderate sample counts, HalfCheetah-v5 reaches teacher-level recovery by 5×10^5 samples, while Ant-v5 and Humanoid-v5 require larger budgets. Humanoid shows the clearest threshold, rising sharply between 2.5×10^5 and 5×10^5 samples.

Implication. The recovery results rule out the simplest interpretations: the student is not learning from trajectories, rewards, environment interaction, or physically valid states. The remaining explanation is geometric: recovery depends on covering the directions and regions of input space that control the teacher's action function. The next section develops this interpretation through a sensitivity-overlap framework, showing that directional alignment and ambient support jointly govern recovery. Smoothness alone, however, does not predict recovery success (Appendix D.6).

5 Geometry of Policy Recovery

The previous section establishes the empirical puzzle: physically invalid synthetic observations can recover strong policies, but not all synthetic distributions work. We now test a geometric explanation. Recovery depends on two distinct properties of the query distribution: it must excite directions along which the teacher changes, and it must retain sufficient variance in deployment-relevant directions outside the dominant sensitivity subspace.

5.1 SENSITIVITY OVERLAP AND COVERAGE

We measure directional excitation using the teacher sensitivity matrix

$$G_{\pi_T} = \mathbb{E}_{x\sim\rho_{\text{eval}}}\left[J_{\pi_T}(x)^\top J_{\pi_T}(x)\right], \tag{1}$$

where $J_{\pi_T}(x)$ is the teacher Jacobian evaluated on real rollout observations. For a zero-mean query distribution q with covariance Σ_q, we define the raw sensitivity-overlap score

$$\mathcal{C}(q) = \text{Tr}(G_{\pi_T}\Sigma_q). \tag{2}$$

For reporting across interventions, we additionally use the normalized score

$$\bar{\mathcal{C}}(q) = \frac{\mathcal{C}(q)}{\text{Tr}(G_{\pi_T})}, \tag{3}$$

so that $\bar{\mathcal{C}}(q) = 1$ for the unit isotropic Gaussian $q = \mathcal{N}(0, I)$. If the teacher is sensitive along direction v_j, then $v_j^\top \Sigma_q v_j$ measures how strongly q excites that direction. When support dimension and per-direction variance are matched, $\mathcal{C}(q)$ therefore measures how query energy is allocated across teacher-sensitive directions. We call $\mathcal{C}(q)$ a *raw* overlap score because neither it nor its normalized version is intended to rank distributions with different total variance or ambient support.

Directional overlap alone cannot guarantee deployment accuracy. Let q and ρ denote the query and deployment distributions, with second-moment matrices Σ_q and Σ_ρ. For a local linear student–teacher error $e(x) = Ax$, define $L_q(A) = \text{Tr}(A\Sigma_q A^\top)$ and analogously $L_\rho(A)$. If $\Sigma_q \succ 0$, then

$$L_\rho(A) \le \underbrace{\lambda_{\max}\left(\Sigma_q^{-1/2}\Sigma_\rho\Sigma_q^{-1/2}\right)}_{\text{coverage mismatch}} L_q(A). \tag{4}$$

If Σ_q is singular and deployment has variance in its null space, two policies can agree on every query while differing arbitrarily at deployment. Strict subspace matching therefore gives no general deployment guarantee. Equation 4 is our primary theoretical coverage statement.

To illustrate the corresponding finite-sample effect, consider a Gaussian linear-regression surrogate with effective residual variance σ^2. For $\Sigma_q \succ 0$ and $N > d + 1$,

$$\mathbb{E}L_\rho(\widehat{A} - A) = \frac{m\sigma^2}{N - d - 1}\underbrace{\text{Tr}\left(\Sigma_\rho\Sigma_q^{-1}\right)}_{\mathcal{R}_{\text{inv}}}. \tag{5}$$

This surrogate is not intended as an exact model of neural student training: teacher labels are deterministic, and σ^2 represents effective residual error arising from local nonlinear approximation, optimization, or model misspecification. We therefore use $\mathcal{R}_{\text{inv}}$ as a coverage diagnostic motivated by the finite-sample analysis, rather than as a universal predictor of nonlinear recovery. Full proofs are given in Appendix G.

5.2 DIRECTIONAL COVERAGE THROUGH SUBSPACE INTERVENTION

We summarize the teacher sensitivity spectrum using r_{90}, the number of directions explaining 90% of total sensitivity, and the participation-ratio dimension

$$d_{\text{sens}} = \frac{(\sum_j \lambda_j)^2}{\sum_j \lambda_j^2}.$$

Synthetic probes closely recover the dominant teacher-sensitive subspace, but this observation is correlational; full spectra and real-versus-synthetic alignment are reported in Appendix D.3. We therefore intervene directly on the sampled directions.

Let $V_k \in \mathbb{R}^{d\times k}$ be an orthonormal basis, sample $z \sim \mathcal{N}(0, I_k)$, and set $x = V_k z$. This fixes support dimension and per-active-direction variance while changing only the sampled subspace. At a budget of 10^5 queries and 20 epochs, we compare Full Gaussian, Top-k, Random-k, and Bottom-k sampling for $k = \lceil d_{\text{sens}} \rceil$ and $k = r_{90}$.

Table 1: Subspace intervention recovery on Humanoid-v5 and Ant-v5 (mean $\pm$ std, 10 seeds). Normalized sensitivity-overlap scores are shown in parentheses.

Condition	Humanoid (%)	Ant (%)
Full Gaussian (1.000)	39.8 ± 18.0	81.0 ± 11.2
Top-d_{sens} (0.810 / 0.736)	16.4 ± 9.9	25.8 ± 13.1
Random-d_{sens} (0.046 / 0.067)	3.6 ± 1.8	18.7 ± 2.5
Bottom-d_{sens} (0.003 / 0.002)	2.8 ± 0.8	16.9 ± 4.3
Top-r_{90} (0.901 / 0.897)	25.0 ± 9.1	39.6 ± 19.3
Random-r_{90} (0.224 / 0.162)	4.5 ± 1.3	16.5 ± 3.7

Table 2: Thickened Top-78 recovery on Humanoid-v5 (mean $\pm$ std across 10 matched teachers, 500k samples, 100 epochs). $\varepsilon = 0$ is strict Top-78 and $\varepsilon = 1$ approximates Full Gaussian.

ε	Overlap / $\mathrm{Tr}(G)$	Recovery (%)
0	0.901 ± 0.014	8.1 ± 2.5
0.05	0.899 ± 0.015	56.8 ± 19.4
0.10	0.900 ± 0.015	73.7 ± 19.7
0.25	0.905 ± 0.014	71.7 ± 14.0
0.50	0.924 ± 0.011	73.3 ± 17.4
1.00	1.000 ± 0.000	63.5 ± 19.3

The intervention confirms the directional effect. On Humanoid, Top-16 recovers 16.4%, compared with 3.6% for Random-16 and 2.8% for Bottom-16. At the broader r_{90} scale, Top-78 reaches 25.0% while Random-78 remains at 4.5%. Across the matched intervention conditions, recovery follows the same directional ordering as sensitivity overlap on Humanoid and approximately the same ordering on Ant. However, the remaining gap to Full Gaussian shows that exciting sensitive directions is not sufficient. This supports the use of overlap as a controlled directional diagnostic under matched support and variance, rather than as a general predictor across arbitrary query distributions.

5.3 Ambient support requirement

A strict top subspace provides no supervision in orthogonal directions that may still be encountered during deployment. We isolate this effect by adding controlled orthogonal variance:

$$x = V_{78} z_\top + \varepsilon V_\perp z_\perp, \qquad z_\top \sim \mathcal{N}(0, I_{78}), \quad z_\perp \sim \mathcal{N}(0, I_{d-78}). \tag{6}$$

Adding a small amount of orthogonal variance produces a large recovery gain with almost no change in raw overlap. Table 2 shows this: strict Top-78 recovers only 8.1%, while $\varepsilon = 0.10$ reaches 73.7% as overlap changes from 0.901 to 0.900. Across the same ten teachers, this is a mean paired improvement of 65.6 percentage points (95% CI $[52.0, 79.2]$; paired $t(9) = 10.92$, $p = 1.7 \times 10^{-6}$). An exact two-sided Wilcoxon signed-rank test gives $W = 0$, $p = 0.00195$, yielding the same conclusion without a Gaussian assumption on the paired differences. The strict Top-78 value here is lower than the 25.0% reported in Table 1 because Table 1 uses 100k queries and 20 training epochs, whereas Table 2 uses 500k queries and 100 epochs. Appendix D.5 shows that longer optimization within the restricted subspace can worsen closed-loop recovery despite improving the in-subspace fit.

5.4 Rollout covariance explains the remaining gap

The coverage analysis suggests that isotropic queries may fail because they miss correlated deployment structure rather than simply because they are centered incorrectly. To test this, we use a diagnostic Gaussian fitted to teacher rollout statistics. This experiment uses real rollout data and is therefore not part of the passive synthetic-recovery setting.

Covariance structure accounts for most of the improvement. On Humanoid (Table 3), re-centering alone adds 6.7 percentage points, whereas using rollout covariance adds 34.8 points. The corresponding gains on Ant are 5.0 and 18.9 points. Thus, much of the information missing from $\mathcal{N}(0, I)$ lies

Table 3: Recovery (%) under rollout-informed Gaussian diagnostics.

Query distribution	Humanoid-v5	Ant-v5
$\mathcal{N}(0, I)$	35.9 ± 14.0	78.8 ± 19.9
$\mathcal{N}(\hat{\mu}, I)$	42.5 ± 16.4	83.8 ± 11.7
$\mathcal{N}(0, \hat{\Sigma})$	70.6 ± 19.8	97.6 ± 5.9
$\mathcal{N}(\hat{\mu}, \hat{\Sigma})$	88.7 ± 11.0	99.9 ± 12.7

in correlated deployment structure rather than mean centering alone. This provides an empirical counterpart to the covariance mismatch in Section 5.1. The full factorial decomposition is reported in Appendix D.9.

5.5 Synthetic-only query design

The preceding interventions use rollout-derived eigenspaces. We therefore ask whether useful geometry can instead be estimated without rollout observations. For each teacher, a 1,000-point $\mathcal{N}(0, I)$ pilot provides Jacobians for selecting a top subspace, after which a thickened distribution supplies 99,000 distillation queries. The isotropic baseline receives 100,000 queries, so action-label budgets are matched. No rollout state is used for design or training, although the pilot requires white-box Jacobian access

The pilot-guided design is neutral on Ant and Humanoid and significantly worse on HalfCheetah and Hopper. Increasing retained rank or tail variance repairs most of these failures, but does not reliably outperform isotropic querying. A pilot-size ablation from 500 to 10,000 Jacobian queries does not repair the HalfCheetah or Hopper failures and does not improve the neutral Humanoid result, showing that the observed behavior is not explained by insufficient pilot size (Appendix G.6). Post-design $\mathcal{R}_{\text{inv}}$ strongly predicts action MSE in the two failure environments, whereas raw subspace overlap does not. It is only moderately predictive on Ant and fails on Humanoid. Inverse coverage is therefore a deployment-dependent diagnostic rather than a universal scalar law. Full confidence intervals, action MSEs, ablations, and leave-one-teacher-out tests are in Appendix G.5.

5.6 Breadth and boundary checks

The two-factor effect is not restricted to PPO MuJoCo policies. Table 4 summarizes interventions that change the RL algorithm, action representation, and activation function. In both settings, adding orthogonal support substantially improves recovery while sensitivity overlap remains nearly unchanged. A separate 20-teacher MNIST control provides a boundary: label shuffling raises effective sensitivity rank while synthetic recoverability falls despite nearly unchanged training accuracy, showing that favorable sensitivity geometry is not sufficient for memorization-dominated functions. Full results are in Appendix E.

Table 4: Orthogonal thickening beyond the main PPO locomotion setting.

Setting	Strict top	$+\varepsilon = 0.10$	Overlap
LunarLander, PPO/ReLU	52.0%	89.9%	$0.969 \rightarrow 0.969$
Ant-v5, SAC	-6.0%	71.2%	$0.903 \rightarrow 0.904$

A second boundary separates function recovery from control recovery. On Hopper-v5, increasing the query budget from 250k to 500k reduces action MSE at every tested student width while closed-loop recovery remains near 98–100%, so imitation fidelity can improve after control performance saturates (Appendix F). The modest predictive contribution of d_{sens} does not contradict the subspace interventions: d_{sens} summarizes teacher sensitivity dimension, whereas the interventions change which sensitive directions are covered at matched support dimension and per-direction variance. Thus, directional alignment can causally affect recovery even when d_{sens} is a weak cross-condition predictor.

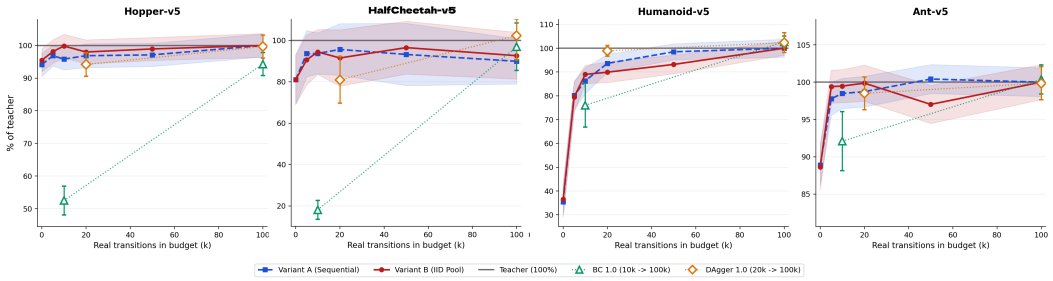

Figure 2: Fixed-budget real-synthetic mixing across four environments. Variant A uses contiguous rollout observations; Variant B samples i.i.d. observations from the same rollout pool. Behavioral cloning (BC) and DAgger use real observations at comparable budgets.

6 Mixed Real-Synthetic Recovery in the Low-Data Regime

The geometry results explain why passive synthetic queries can constrain a policy, but they do not identify the states that matter during closed-loop execution. We therefore ask whether a small rollout buffer can supply this missing deployment information. Every student receives exactly 100k teacher labels. We replace 5k–100k synthetic inputs with real teacher-rollout observations and compare contiguous trajectory segments with i.i.d. samples from the same rollout pool.

Figure 2 shows a clear environment-dependent pattern. Ant-v5 and Humanoid-v5 improve sharply after only 5k–20k real inputs and then saturate. Hopper-v5 and HalfCheetah-v5 gain little because their synthetic-only students already recover most teacher performance. Real observations therefore act as rollout-region anchors where passive coverage is insufficient, while the remaining synthetic queries constrain the teacher more broadly.

Contiguous and i.i.d. real samples give statistically indistinguishable recovery across environments and budgets. Within this memoryless regression objective, the useful signal comes mainly from the visited-state marginal, not from temporal ordering. This comparison does not remove the student-induced distribution shift addressed by interactive methods such as DAgger. It also does not show that the synthetic distribution becomes localized around real states, because the synthetic component remains unchanged. Thus, passive queries provide broad functional constraints, while small real buffers anchor the student to deployment-relevant states when synthetic recovery is incomplete.

7 Conclusion

We show that trained state-based reinforcement learning policies can recover substantial closed-loop performance from independently sampled, teacher-labeled synthetic inputs without trajectories, rewards, or environment interaction. Recovery is governed by two complementary forms of geometric coverage: queries must excite teacher-sensitive directions while retaining sufficient support outside the dominant sensitivity subspace. Matched subspace interventions and orthogonal thickening isolate these effects empirically. The theoretical results provide the corresponding explanation: strict subspace queries can leave deployment behavior unidentifiable, while covariance mismatch controls how query-space error transfers to deployment. Small real-state buffers provide a complementary benefit by anchoring the student to the visited-state distribution.

These results also identify important boundaries. Teacher sensitivity can be estimated from synthetic queries alone, but sensitivity-guided query design does not consistently outperform isotropic sampling, which remains a strong baseline when deployment geometry is unknown. Moreover, improved action imitation need not imply improved control performance: action error can continue to decrease after return has saturated. Our conclusions are therefore limited to the compact vector-input policies studied here and should not be interpreted as universal recovery guarantees across architectures or observation modalities. Overall, the results characterize when passive synthetic queries can recover useful policy behavior and why geometric coverage, rather than physical validity alone, is the central requirement.

Reproducibility statement

Appendices D.1–E report teacher and student architectures, training budgets, sampling distributions, seed counts, evaluation protocols, full recovery tables, intervention definitions, and statistical tests. Appendix F reports the additional scaling design and its negative transfer results. Appendix G gives the coverage proofs and full synthetic-pilot results. The supplementary source package contains the complete manuscript inputs and figure assets.

AI use statement

Generative AI tools assisted with experimental-design feedback, code drafting and formatting. The authors executed the experiments, checked reported values against experimental outputs, reviewed all AI-assisted material, and take responsibility for the final text, claims, code, and artifacts.

References

Josh Achiam and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning, 2019. URL https://api.semanticscholar.org/CorpusID:208283920.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3513–3521. IEEE, 2019.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Chunfeng Cui, Kaiqi Zhang, Talgat Daulbaev, Julia Gusak, Ivan Oseledets, and Zheng Zhang. Active subspace of neural networks: Structural analysis and universal attacks. *SIAM Journal on Mathematics of Data Science*, 2(4):1096–1122, 2020.

Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1331–1340. PMLR, 2019.

Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Alexandre Galashov, Josh S Merel, and Nicolas Heess. Data augmentation for efficient learning from parametric experts. *Advances in Neural Information Processing Systems*, 35:31484–31496, 2022.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. Demodice: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2021.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

Andrei Lupu, Chris Lu, Jarek Liesen, Robert Lange, and Jakob Foerster. Behaviour distillation. In *International Conference on Learning Representations*, volume 2024, pp. 12668–12684, 2024.

Yecheng Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In *International Conference on Machine Learning*, pp. 14639–14663. PMLR, 2022.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *International conference on machine learning*, pp. 4743–4751. PMLR, 2019.

Gaurav Kumar Nayak, Konda Reddy Mopuri, and Anirban Chakraborty. Effectiveness of arbitrary transfer sets for data-free knowledge distillation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1429–1437. IEEE, 2021.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. corr abs/1812.02766 (2018). *arXiv preprint arXiv:1812.02766*, 2018.

Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.

Piyush Raikwar and Deepak Mishra. Discovering and overcoming limitations of noise-engineered data-free knowledge distillation. *Advances in Neural Information Processing Systems*, 35:4902–4912, 2022.

Jonathan Raiman. Generative adversarial simulator. *arXiv preprint arXiv:2011.11472*, 2020.

Juntao Ren, Gokul Swamy, Zhiwei Steven Wu, J Andrew Bagnell, and Sanjiban Choudhury. Hybrid inverse reinforcement learning. *arXiv preprint arXiv:2402.08848*, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4769–4778. IEEE, 2021.

Haoran Xu, Xianyuan Zhan, Honglei Yin, and Huiling Qin. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *International Conference on Machine Learning*, pp. 24725–24742. PMLR, 2022.

Hongxu Yin, Pavlo Molchanov, José Manuel Álvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj Kumar Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8712–8721, 2019. URL `https://api.semanticscholar.org/CorpusID:209405263`.

A Limitations and Scope

Our empirical scope is limited to compact vector-input policies. The main experiments use PPO-family MuJoCo actors, with targeted checks on SAC, discrete-action ReLU policies, and constrained PPO-Lag teachers. These experiments do not establish algorithm- or architecture-universal recovery. In particular, the same passive synthetic-query procedure does not transfer directly to pixel-based policies, and recurrent, Transformer, graph-structured, and multi-agent policies remain untested. Our conclusions should therefore be restricted to state-based policies of the type studied here.

The geometric analysis also has important limitations. Most sensitivity quantities are computed post-hoc from teacher rollout observations, and although the synthetic-only pilot estimates teacher geometry without rollout data, its query design does not reliably outperform isotropic sampling. Moreover, inverse-coverage diagnostics require deployment covariance after query design, and the normalized policy-input space itself provides a coordinate-scale prior. Thus, the present results characterize when passive synthetic recovery succeeds rather than providing a fully data-free procedure for constructing an optimal query distribution. Synthetic distillation also does not create robustness, safety, or dynamics knowledge absent from the teacher, and students may inherit the teacher's failures.

B Behavioral Fidelity Under Perturbations and Constraints

The previous sections show that synthetic observations can recover the teacher's policy structure under nominal dynamics. We now ask whether students also preserve the teacher's response under off-nominal evaluation. This is not a test of whether synthetic distillation creates robustness beyond the teacher: the student receives no reward signal, environment interaction, domain randomization, or robustness objective. The relevant question is whether the student tracks the teacher's perturbation-response profile.

We evaluate teacher, behavioral cloning, DAgger, and synthetic students under additive observation noise, gravity scaling, and body-mass scaling. Policies are evaluated without retraining, and returns are normalized by each method's own unperturbed return.

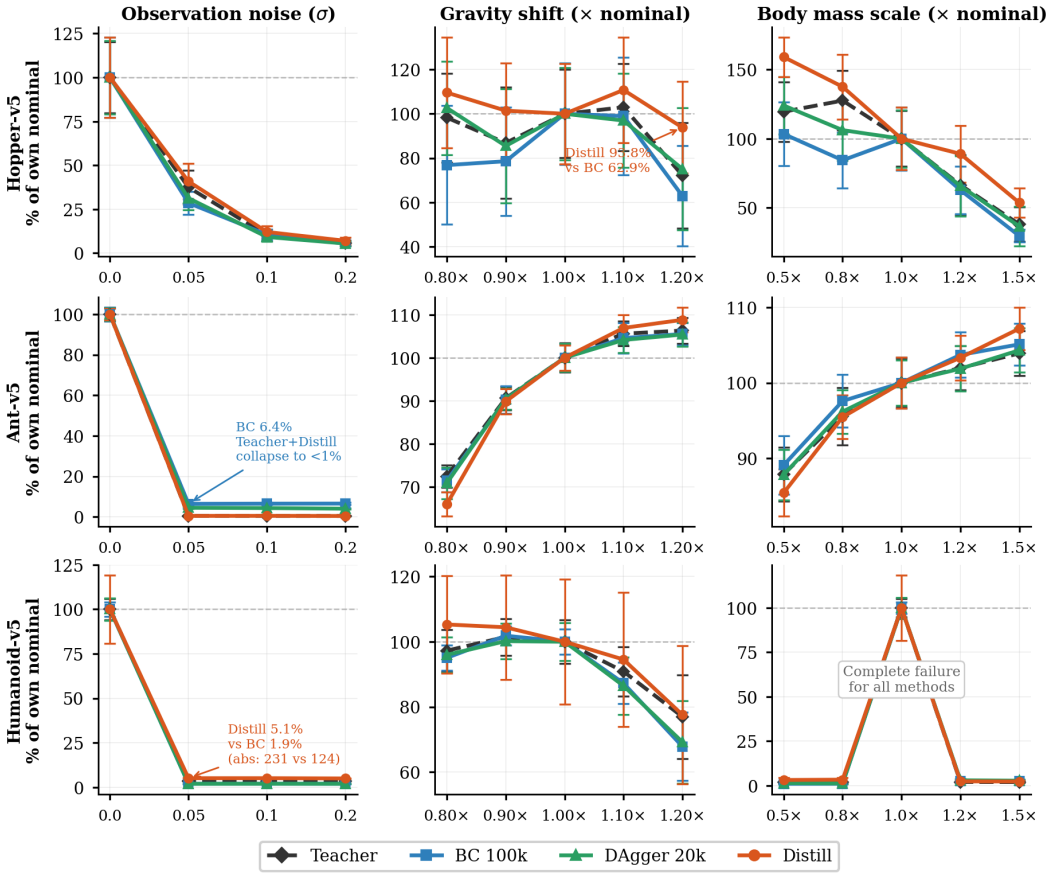

Figure 3: Perturbation fidelity under observation and dynamics shifts. Returns are normalized by each method's nominal return. Synthetic students generally inherit the teacher's perturbation profile: they degrade when the teacher degrades and often retain performance when the teacher does. The goal is behavioral fidelity, not robustness improvement.

Figure 3 shows that synthetic students largely inherit the teacher's perturbation profile. Observation noise gives the clearest case: across Hopper-v5, Humanoid-v5, and Ant-v5, the teacher degrades sharply as noise increases, and synthetic students degrade in the same regime. This is expected. The student has no mechanism for learning invariance to arbitrary observation corruption unless such invariance is already encoded in the teacher's action map and covered by the distillation distribution.

Dynamics perturbations are more environment-dependent. In Hopper-v5, synthetic students retain more nominal-normalized performance than behavioral cloning under several gravity and mass shifts. In Humanoid-v5, gravity shifts degrade all methods similarly, and body-mass scaling causes

near-complete failure except near nominal mass. Ant-v5 remains comparatively stable under gravity and mass shifts but collapses under observation noise.

These results separate policy-function fidelity from robustness improvement. Synthetic recovery can match the teacher's observation-to-action behavior well enough that downstream rollout behavior co-varies with the teacher under perturbation. It does not learn new invariances or control strategies absent from the teacher. In this sense, the student inherits the teacher's behavioral geometry, including both strengths and failures. Appendix C gives the complementary constrained-control case: when the teacher is trained with a reward-cost objective, synthetic students can approximately preserve its empirical reward-cost behavior.

C SYNTHETIC RECOVERY OF A SAFETY-AWARE TEACHER

The main experiments study reward-only MuJoCo locomotion policies. To test whether synthetic recovery is limited to this setting, we include a breadth check on a constrained-control domain from Safety-Gymnasium. Safety-Gymnasium provides safety-critical reinforcement-learning environments with both vector and vision inputs, together with OmniSafe, a library of safe-RL algorithms. We use SafetyPointGoal1, a compact vector-observation navigation task in which the agent must reach a goal while avoiding hazards. Unlike the MuJoCo locomotion policies in the main text, the teacher is trained with PPO-Lag, a Lagrangian variant of PPO that controls constraint violations through a learned Lagrange multiplier (Achiam & Amodei, 2019).

This experiment is not intended to evaluate synthetic distillation as a safe-RL algorithm. The student is not trained with rewards, costs, constraints, or environment interaction. As in the main experiments, the student receives only independently sampled synthetic observations labeled by the fixed teacher policy. The purpose is narrower: to test whether synthetic recovery persists when the teacher was trained under a reward-cost objective and the observation contains explicit safety-relevant structure.

Setup. We train PPO-Lag teachers on SafetyPointGoal1 across 10 seeds and distill students from teacher-labeled synthetic observations using the same supervised regression objective as Section 3. Each student is evaluated by executing the learned policy in the original SafetyPointGoal1 environment. We report episodic reward, episodic cost, and empirical Lipschitz estimates as mean $\pm$ standard deviation across the 10 teacher seeds. Reward recovery is computed relative to the corresponding PPO-Lag teacher Cost is reported directly and as a difference from the teacher: $\Delta C = C(\pi_S) - C(\pi_T)$. Negative values indicate lower episodic cost than the teacher. For reward recovery and cost difference, we compute the paired quantity within each seed before averaging across seeds.

Aggregate recovery. Table 5 reports aggregate performance across all synthetic distributions, sample budgets, widths, and seeds. For the student row, each seed is first averaged over all synthetic configurations, and the reported standard deviation is then computed across all seeds. Synthetic students recover nearly all teacher reward while maintaining lower episodic cost on average. This shows that the recovery phenomenon is not restricted to unconstrained MuJoCo locomotion: it also appears for a teacher trained with a constrained reward-cost objective.

Table 5: Synthetic recovery of PPO-Lag teachers on SafetyPointGoal1. Values are mean $\pm$ std across 10 teacher seeds. Reward recovery and Δ cost are computed relative to the corresponding PPO-Lag teacher within each seed. Cost is episodic cost; lower is better.

Policy group	Recovery (%)	Cost	Δ cost
PPO-Lag teacher	100.0	55.70 ± 9.60	0.00
Synthetic students	98.68 ± 1.91	50.33 ± 2.38	-5.37 ± 8.45

Across 10 seeds, PPO-Lag teachers obtain reward 21.72 ± 2.73 and empirical Lipschitz estimate 25.07 ± 2.07. Averaged over all synthetic student configurations, students obtain reward 21.45 ± 2.89 and empirical Lipschitz estimate 37.70 ± 2.48. Thus, the average synthetic student recovers 98.68% of teacher reward while reducing mean episodic cost by 5.37 relative to the teacher. This is a strong recovery result because the student has no access to costs, constraints, rewards, transitions, or valid trajectories during training. The only supervision is the teacher's action on independently sampled synthetic observations.

These results suggest that the learned PPO-Lag policy contains recoverable structure in its observation-to-action map, just as the MuJoCo teachers do. The result is especially relevant because the teacher's behavior depends on hazard-related observation channels, yet synthetic supervision approximately preserves both task performance and empirical rollout cost.

At the same time, this experiment should not be interpreted as a safe-RL algorithm. The student does not optimize a constraint objective, does not observe costs, and does not receive safety feedback. Instead, it can preserve favorable reward-cost behavior only insofar as that behavior is encoded in the teacher's action function and recoverable from synthetic queries. Thus, the experiment provides breadth evidence for synthetic policy recovery, not a claim that synthetic distillation enforces safety constraints by itself.

D Experimental Details

This appendix documents hyperparameters, seed counts, compute resources, and full numerical results underlying the figures and tables in the main paper. It also contains supporting analyses of smoothness (D.6), teacher–student sensitivity alignment (D.8, D.4), student sensitivity on synthetic versus real inputs (D.10), and the concentration-of-measure effect behind uniform-wide sampling in high-dimensional environments (D.12).

D.1 Hyperparameters

All four environments use Stable-Baselines3 PPO with the algorithm defaults shown below. The teacher is a shared-trunk MLP with separate actor and critic heads. Both heads use orthogonal initialization with std $= \sqrt{2}$ for hidden layers, std $= 0.01$ for the actor output, and std $= 1.0$ for the critic output.

Table 6: PPO hyperparameters per environment. Entries not listed are SB3 defaults.

Hyperparameter	Hopper-v5	HalfCheetah-v5	Ant-v5	Humanoid-v5
Clip range	0.2	0.2	0.2	0.2
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
n_{steps} (rollout)	2,048	2,048	2,048	1,024
PPO epochs	10	10	10	10
Batch size	256	256	4,096	2,048
Entropy coefficient	0.0	0.0	0.0	0.0
Value coefficient	0.5	0.5	0.5	0.5
GAE λ	0.95	0.95	0.95	0.95
γ (discount)	0.99	0.99	0.99	0.99
Hidden sizes	$[64, 64]$	$[64, 64]$	$[256, 256]$	$[512, 512]$
Activation	Tanh	Tanh	Tanh	Tanh
Total timesteps	2×10^6	2×10^6	1×10^7	1×10^7
Parallel envs	4	4	16	32
Obs. normalization	on	on	on	on
Reward scaling	on	on	on	on
Eval freq. (steps)	10^5	10^5	10^5	10^5

Observation and reward normalization are enabled through the stored `VecNormalize` statistics used during teacher training and inference. SB3 PPO uses `vf_coef` $= 0.5$ and `max_grad_norm` $= 0.5$; neither is overridden.

Students are trained by supervised regression of the teacher's deterministic action mean. The loss is mean-squared error on $(\pi_S(o), \pi_T(o))$ over the synthetic dataset, with no auxiliary terms.

For all main-text mixing experiments and for all appendix tables in this document, only the full-width ($1.0\times$) student is reported. The reduced-width variants are part of a separate ablation and do not affect any conclusion in the main paper. All experiments were run on a single Apple M3 Max (CPU/MPS, no discrete GPU) using PyTorch 2.4.1.

Table 7: Student distillation hyperparameters.

Hyperparameter	Value
Optimizer	Adam (default $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	3×10^{-4}
Batch size	2,048
Training epochs	100
Loss function	MSE on deterministic action mean
Gradient clipping	$\|\nabla\|_2 \leq 0.5$
Architecture	2-layer MLP; width $\in \{1.0\times, 0.5\times, 0.25\times\}$ teacher width
Initialization	Orthogonal (std $= \sqrt{2}$ hidden, 0.01 output)
Activation	Tanh
Obs. normalization	on
Sample budgets	$\{10{,}000, 25{,}000, 50{,}000, 100{,}000, 250{,}000, 500{,}000, 1{,}000{,}000, 2{,}000{,}000\}$

Synthetic observations are sampled in normalized observation space. The teacher's stored running statistics $(\mu_{\text{rms}}, \sigma_{\text{rms}})$ from `VecNormalize` are used to map between raw and normalized observations: synthetic samples are generated as normalized vectors and consumed directly by the policy network; during evaluation rollouts in the nominal environment, raw observations are mapped to normalized space via $x = (o - \mu_{\text{rms}})/\sigma_{\text{rms}}$ before policy inference.

D.2 Full numerical results for Figure 1

This subsection reports the underlying recovery percentages for the synthetic distillation curves in Figure 1. Values are normalized to the teacher's mean return (not to any other condition); the teacher's per-environment return is shown in each subtable header. Entries are mean $\pm$ std across 10 student seeds trained on full-width architectures. The tables below report full-width recovery curves. Hopper-v5 saturates near 95% at the 100k budget. HalfCheetah-v5 is shown through 500k samples, where Gaussian, Laplace, and uniform students reach teacher-level recovery. Figure 1 in the main text additionally shows recovery up to 2M samples for Ant-v5 and Humanoid-v5 from upscaled checkpoints.

Table 8: Hopper-v5 recovery (%, mean $\pm$ std) across 10 seeds. Teacher mean return $= 1706 \pm 617$.

Distribution	10k	25k	50k	100k
Gaussian, $\sigma = 1$	38.5 ± 9.4	81.7 ± 14.5	84.1 ± 14.8	93.7 ± 8.5
Gaussian, $\sigma = 2$ (wide)	44.6 ± 10.8	77.2 ± 23.6	81.3 ± 17.5	88.8 ± 15.5
Gaussian, $\sigma = 0.5$ (narrow)	33.8 ± 11.0	59.7 ± 15.9	73.2 ± 16.0	79.2 ± 17.4
Gaussian, $\mu = 1$, $\sigma = 1$ (shifted)	22.8 ± 4.3	46.2 ± 11.1	63.7 ± 13.7	89.4 ± 26.9
Uniform $[-1, 1]$ (wide)	33.2 ± 8.7	65.1 ± 22.5	78.7 ± 7.6	85.8 ± 16.8
Uniform $[-0.5, 0.5]$ (narrow)	29.4 ± 12.6	34.4 ± 8.4	46.9 ± 16.0	66.6 ± 20.1
Laplace, $b = 1$	45.8 ± 11.5	79.4 ± 20.2	86.4 ± 10.5	96.9 ± 15.7

Table 9: HalfCheetah-v5 recovery (%, mean $\pm$ std) across 10 seeds. Teacher mean return $= 3181 \pm 1216$.

Distribution	10k	25k	50k	100k	250k	500k
$\mathcal{N}(0, 1)$	13.2 ± 5.2	39.0 ± 16.5	61.6 ± 28.9	83.5 ± 33.6	107.4 ± 38.4	108.6 ± 40.0
$\mathcal{N}(0, 2)$	13.8 ± 5.5	29.5 ± 14.4	49.7 ± 26.3	69.5 ± 34.5	109.1 ± 38.9	110.6 ± 39.6
$\mathcal{N}(0, 0.5)$	12.6 ± 4.9	17.1 ± 7.3	46.8 ± 25.0	65.7 ± 43.9	100.6 ± 39.9	99.9 ± 38.5
$\mathcal{N}(1, 1)$	8.3 ± 8.3	11.3 ± 3.5	18.2 ± 7.6	30.8 ± 13.5	68.8 ± 47.0	89.5 ± 45.1
$\mathcal{U}[-1, 1]$	13.2 ± 5.2	29.4 ± 17.2	46.3 ± 26.3	60.6 ± 40.2	101.8 ± 41.9	108.9 ± 39.7
$\mathcal{U}[-0.5, 0.5]$	11.2 ± 4.3	13.3 ± 5.5	14.9 ± 5.6	27.4 ± 12.7	89.4 ± 51.4	96.8 ± 33.9
$\mathcal{L}(0, 1)$	13.4 ± 4.7	31.7 ± 15.7	51.6 ± 17.6	83.9 ± 37.8	106.3 ± 41.0	108.1 ± 41.1

These are pure synthetic distillation results without real-data mixing; mixing results are reported in Appendix D.11.

Table 10: Ant-v5 recovery (%, mean ± std) across 10 seeds. Teacher mean return = 4938 ± 957.

Distribution	10k	25k	50k	100k
$\mathcal{N}(0,1)$	46.4 ± 20.3	57.0 ± 20.2	72.8 ± 15.4	81.9 ± 9.6
$\mathcal{N}(0,2)$	51.9 ± 14.2	55.4 ± 19.2	64.5 ± 19.2	78.8 ± 13.3
$\mathcal{N}(0,0.5)$	36.2 ± 21.3	45.3 ± 21.7	61.0 ± 18.8	77.6 ± 14.9
$\mathcal{N}(1,1)$	33.6 ± 17.6	34.5 ± 18.0	37.2 ± 22.7	39.2 ± 22.4
$\mathcal{U}[-1,1]$	38.0 ± 21.9	47.8 ± 23.4	71.9 ± 13.1	79.5 ± 6.8
$\mathcal{U}[-0.5,0.5]$	21.5 ± 15.4	25.2 ± 20.8	42.6 ± 18.5	62.9 ± 16.7
$\mathcal{L}(0,1)$	49.7 ± 16.7	59.4 ± 17.6	70.0 ± 13.8	80.8 ± 11.0

Table 11: Humanoid-v5 recovery (%, mean ± std) across 10 seeds. Teacher mean return = 6673.42 ± 757.35.

Distribution	10k	25k	50k	100k	250k	500k	1M	2M
$\mathcal{N}(0,1)$	34.5 ± 8.6	35.0 ± 10.2	35.5 ± 11.9	36.4 ± 11.5	44.3 ± 17.2	67.9 ± 23.0	69.8 ± 24.3	77.9 ± 18.9
$\mathcal{N}(0,2)$	30.3 ± 6.5	27.1 ± 5.4	27.4 ± 8.2	31.6 ± 8.2	36.6 ± 15.3	47.4 ± 25.2	67.4 ± 26.6	69.5 ± 29.2
$\mathcal{N}(0,0.5)$	37.0 ± 13.0	38.1 ± 13.3	40.8 ± 13.3	45.0 ± 15.2	49.1 ± 23.6	64.7 ± 27.5	71.6 ± 26.6	72.0 ± 23.3
$\mathcal{N}(1,1)$	22.5 ± 13.4	23.2 ± 14.7	24.8 ± 13.4	27.8 ± 15.4	33.9 ± 19.6	48.1 ± 25.4	38.2 ± 22.4	47.2 ± 26.2
$\mathcal{U}[-1,1]$	34.7 ± 10.1	38.8 ± 12.3	39.1 ± 11.2	44.9 ± 15.2	47.5 ± 22.6	64.4 ± 25.1	69.1 ± 24.9	74.3 ± 20.9
$\mathcal{U}[-0.5,0.5]$	40.3 ± 14.8	40.4 ± 17.7	42.6 ± 15.3	43.2 ± 15.1	47.6 ± 21.1	51.3 ± 24.6	64.3 ± 25.9	63.5 ± 25.8
$\mathcal{L}(0,1)$	35.8 ± 11.4	30.3 ± 7.6	30.0 ± 7.6	33.9 ± 7.1	40.3 ± 18.7	62.1 ± 28.2	72.1 ± 26.7	77.3 ± 22.0

The coordinate-wise shifted Gaussian is a deliberately off-centered stress test, but it also increases aggregate input norm in high-dimensional spaces. Table 12 controls for this by shifting the mean along a random direction while keeping the expected squared norm equal to that of $x \sim \mathcal{N}(0, I_d)$. The milder $c = 0.50$ shift often matches the centered Gaussian baseline, indicating that small off-center shifts are not necessarily harmful. In contrast, the larger $c = 0.75$ shift consistently underperforms the centered Gaussian baseline across HalfCheetah-v5, Ant-v5, and Humanoid-v5. Thus, shifted-distribution failure is not solely attributable to larger aggregate input norm or activation saturation; distributional placement still matters.

D.3 Sensitivity Structure Under Synthetic Probing

This section reports the full sensitivity-structure analysis underlying Table 13. We compare the input-sensitivity geometry of the teacher policy as measured on real rollout observations versus synthetic inputs drawn from various distributions in normalized observation space.

Setup. For each teacher policy we compute $G_\pi = \frac{1}{n}\sum_i J_i^\top J_i$, where $J_i = \partial\pi(x_i)/\partial x_i$ is evaluated at inputs x_i. For real observations, x_i are sampled from rollout trajectories. For synthetic probing, $x_i \sim q(x)$ where q is a synthetic distribution in normalized observation space. All Jacobians use autograd on the normalized actor network; 1000 probe points per seed for the main tables, 10000 for the robustness check below.

Humanoid real-versus-synthetic sensitivity structure. Table 13 gives the detailed Humanoid-v5 comparison that motivates the controlled interventions in the main text. The dominant sensitivity subspace is highly stable across centered synthetic probe distributions even though those probes are not valid rollout states.

At the dominant scale ($k = 16$), centered synthetic distributions align almost identically with the rollout-derived teacher eigenspace ($\cos\bar{\theta} \approx 0.988$), while shifted distributions align less well (0.949–0.955). At the broader r_{90} scale, the centered distributions converge to a common alignment near 0.76, indicating that simple distributional shape does not recover the full sensitivity tail. The real-fit Gaussian analysis in Appendix D.9 shows that off-diagonal rollout covariance explains much of the remaining gap. These measurements are descriptive; the matched subspace and thickening interventions in Section 5 provide the controlled tests.

Per-environment teacher sensitivity spectrum. Table 14 reports the teacher-side sensitivity spectrum across all four environments. The $d_{\text{sens}}/d_{\text{obs}}$ ratio decreases sharply with ambient dimension, confirming that higher-dimensional environments are more anisotropic in relative terms.

Table 12: Norm-controlled random-shift control. Values are recovery percentages, mean ± std across 10 seeds. All distributions are defined in normalized policy-input space. The norm-controlled distribution samples $x \sim \mathcal{N}(\alpha u, \sigma^2 I_d)$ with $\alpha = c\sqrt{d}$ and $\sigma = \sqrt{1-c^2}$, where u is a random unit direction, so that $\mathbb{E}\|x\|_2^2 = d$, matching the standard Gaussian baseline.

Environment	Budget	Gaussian	Original shifted	Norm-shift $c = 0.50$	Norm-shift $c = 0.75$
HalfCheetah-v5	100k	88.3 ± 17.1	37.7 ± 20.3	86.7 ± 26.9	36.2 ± 22.5
HalfCheetah-v5	500k	98.2 ± 6.6	85.5 ± 40.8	99.7 ± 11.0	81.9 ± 27.5
Ant-v5	100k	81.9 ± 8.3	40.0 ± 24.0	66.8 ± 19.4	34.2 ± 17.9
Ant-v5	2M	94.0 ± 3.9	69.3 ± 14.1	98.0 ± 7.7	72.2 ± 18.0
Humanoid-v5	100k	35.8 ± 8.6	27.1 ± 13.1	38.5 ± 11.7	28.2 ± 17.5
Humanoid-v5	2M	80.9 ± 12.8	48.7 ± 23.9	77.7 ± 17.5	70.9 ± 22.2

Table 13: Sensitivity structure under real and synthetic inputs on Humanoid-v5 ($d_{\text{obs}} = 348$, 10 seeds). $\cos\bar{\theta}$ measures alignment with the real-observation teacher eigenspace at the dominant top-k subspace ($k = 16$) and the broader r_{90} subspace ($k = 78$).

Source	d_{sens}	$\cos\bar{\theta}$ (top-k)	$\cos\bar{\theta}$ (r_{90})
Real	15.6 ± 1.2	1.000	1.000
Gaussian	17.5 ± 0.8	0.988 ± 0.003	0.764 ± 0.034
Uniform	12.7 ± 0.6	0.988 ± 0.003	0.766 ± 0.032
Scale mixture	15.0 ± 0.7	0.988 ± 0.003	0.764 ± 0.032
Gaussian-unif. mix.	14.0 ± 0.6	0.988 ± 0.002	0.767 ± 0.033
Shifted	21.5 ± 1.1	0.949 ± 0.007	0.724 ± 0.033
Norm-shift ($c = 0.75$)	15.7 ± 0.7	0.955 ± 0.005	0.729 ± 0.030

Cross-environment real vs. synthetic alignment. Table 15 extends the two-column alignment analysis of Table 13 across all four environments. The $k = r_{90}$ alignment floor is an environment-level property: it varies from ≈ 0.98 on Hopper to ≈ 0.76 on Humanoid, and within each environment no centered distribution family escapes it.

D.4 Per-environment subspace overlap

Top-k overlap at $k = \lceil d_{\text{sens}} \rceil$. Table 16 reports teacher–student subspace overlap at the participation-ratio dimension across environments. Overlap is high in all cases. For Hopper, where d_{sens} nearly equals the ambient dimension, this is essentially trivial. For HalfCheetah, Ant, and Humanoid, high overlap provides evidence that students recover the teacher's dominant sensitivity geometry rather than a low-dimensional approximation aligned by chance.

Broader subspace at $k = r_{90}$. Table 17 reports alignment at the r_{90} subspace for Humanoid-v5, together with reference baselines needed to interpret the $\cos\bar{\theta} \approx 0.76$ value in Table 13.

The random-subspace baseline (0.411) gives the expected alignment of a randomly oriented k-dimensional subspace of the same dimension in a 348-dimensional space. The teacher self-consistency ceiling (0.929) gives the alignment between two independent estimates of the teacher's r_{90} eigenspace from split-half rollout observations, representing the highest alignment achievable given finite-sample estimation noise. The student alignment (0.598 unweighted, 0.520 variance-weighted) sits 36% of the way from random to ceiling at 10^5 samples, indicating partial but incomplete recovery of the sensitivity tail.

The variance-weighted overlap (0.520) weights each eigendirection by its teacher eigenvalue λ_T, giving more weight to directions that carry more sensitivity. That it falls modestly below the unweighted measure (0.598) indicates the alignment gap is distributed across the subspace rather than concentrated in low-eigenvalue tail directions. The eigenvalue ratio $\lambda_{16}/\lambda_{78} \approx 22$ confirms that top-16 directions carry substantially more sensitivity than ranks 17–78, but the r_{90} subspace is itself stably identifiable from finite samples (self-consistency 0.929), ruling out estimation noise as the explanation for the gap.

Table 14: Per-environment teacher sensitivity spectrum. r_{90} and r_{95} are the numbers of eigendirections capturing 90% and 95% of total Jacobian variance; d_{sens} is the participation-ratio sensitivity dimension. Mean $\pm$ std across 10 seeds. All values use normalized observations and autograd Jacobians.

Environment	d_{obs}	r_{90}	r_{95}	d_{sens}	$d_{\text{sens}}/d_{\text{obs}}$
Hopper-v5	11	4.2 ± 0.6	5.5 ± 0.7	3.46 ± 0.53	0.315
HalfCheetah-v5	17	9.0 ± 0.7	11.7 ± 0.7	5.91 ± 0.80	0.348
Ant-v5	105	15.4 ± 2.0	30.4 ± 3.7	8.29 ± 0.73	0.079
Humanoid-v5	348	77.8 ± 21.6	168.3 ± 22.4	15.50 ± 1.10	0.045

Table 15: Sensitivity structure under real and synthetic inputs across environments. $\cos\bar{\theta}$ measures alignment with the real-observation teacher eigenspace at two scales: $k = \lceil d_{\text{sens}}(\text{real}) \rceil$ and $k = r_{90}(\text{real})$. Mean $\pm$ std across 10 seeds.

Env.	Source	d_{sens}	$\cos\bar{\theta}$ (top-k)	$\cos\bar{\theta}$ (r_{90})
Hopper	Real	3.46 ± 0.53	1.000	1.000
	Gaussian	3.62 ± 0.43	0.986 ± 0.010	0.984 ± 0.021
	Uniform	3.23 ± 0.32	0.987 ± 0.009	0.984 ± 0.017
	Shifted	3.62 ± 0.70	0.967 ± 0.027	0.960 ± 0.035
HalfCheetah	Real	5.91 ± 0.80	1.000	1.000
	Gaussian	4.90 ± 0.53	0.924 ± 0.035	0.951 ± 0.017
	Uniform	4.39 ± 0.47	0.914 ± 0.043	0.947 ± 0.020
	Shifted	5.02 ± 1.03	0.918 ± 0.030	0.943 ± 0.025
Ant	Real	8.29 ± 0.73	1.000	1.000
	Gaussian	10.41 ± 1.10	0.963 ± 0.009	0.950 ± 0.007
	Uniform	7.43 ± 0.68	0.957 ± 0.015	0.952 ± 0.007
	Shifted	12.50 ± 1.72	0.917 ± 0.016	0.911 ± 0.008
Humanoid	Real	15.60 ± 1.22	1.000	1.000
	Gaussian	17.52 ± 0.80	0.988 ± 0.003	0.764 ± 0.034
	Uniform	12.71 ± 0.55	0.988 ± 0.003	0.766 ± 0.032
	Shifted	21.48 ± 1.11	0.950 ± 0.007	0.724 ± 0.033

D.5 Subspace intervention: full numerical results

Sampling procedure. All sampling occurs in normalized observation space $x = (o - \mu_{\text{rms}})/\sigma_{\text{rms}}$. For each condition we specify an orthonormal basis $V_k \in \mathbb{R}^{d \times k}$ and sample $z \sim \mathcal{N}(0, I_k)$, setting $x = V_k z$. We do not rescale by d/k: per-active-direction variance is held at unity across all conditions so that only the support directions differ. The teacher's empirical sensitivity matrix $G_\pi = n^{-1} \sum_i J_i^\top J_i$ is computed in normalized space using $n = 1000$ real teacher-rollout observations per seed; eigendecomposition gives the top-k, bottom-k, and full eigenvector bases. Random subspaces are sampled by QR-decomposing a random $d \times k$ Gaussian matrix.

Conditions. Humanoid-v5 ($d_{\text{sens}} \approx 16$, $r_{90} \approx 78$): Full $\mathcal{N}(0, I_{348})$, Top-16, Random-16, Bottom-16, Top-78, Random-78. Ant-v5 ($d_{\text{sens}} \approx 7$, $r_{90} \approx 17$): Full $\mathcal{N}(0, I_{105})$, Top-7, Random-7, Bottom-7, Top-17, Random-17. We do not test Bottom-78 or Bottom-17 because in higher-k regimes the bottom subspace contains more variance than the lower-k top subspace and the comparison becomes uninformative.

Training and evaluation. Main-text results use 20 training epochs. We use 20 epochs for the subspace conditions because longer training induces overfitting to the restricted input geometry: students trained strictly within a subspace have no supervision off it, and additional epochs tighten the fit inside the subspace without improving — and often worsening — closed-loop behavior in the full observation space. Table 18 reports results at both 20 and 100 epochs; the directional ordering Top-k $>$ Random-k $>$ Bottom-k is preserved at both epoch counts, confirming that the causal claim does not depend on the epoch choice. Evaluation uses 20 rollouts per student in the nominal environment.

Table 16: Teacher–student subspace overlap at $k = \lceil d_{\text{sens}} \rceil$ across environments. $\cos\bar{\theta}$ is the mean cosine of principal angles. 10 seeds per condition; Gaussian baseline at 10^5 samples.

Environment	$k = \lceil d_{\text{sens}} \rceil$	$\cos\bar{\theta}$
Hopper-v5	4	0.999 ± 0.001
HalfCheetah-v5	6	0.950 ± 0.030
Ant-v5	7	0.982 ± 0.008
Humanoid-v5	16	0.986 ± 0.003

Table 17: Humanoid-v5 subspace overlap at $k = r_{90} = 78$. Teacher self-consistency is computed via split-half estimation on rollout observations from each seed (10 seeds). Random baseline is averaged over 10 random orthonormal bases versus the teacher subspace.

Quantity	$\cos\bar{\theta}$	Interpretation
Random-subspace baseline	0.411 ± 0.001	noise floor
Student vs. teacher (unweighted)	0.598 ± 0.005	observed
Student vs. teacher (variance-weighted)	0.520 ± 0.005	weighted by λ_T
Teacher self-consistency	0.929 ± 0.018	ceiling

D.6 Per-distribution disaggregation of smoothness metrics

This appendix tests whether successful recovery can be explained simply by smoother student functions. It cannot. On Humanoid-v5, shifted-distribution students are smoother than Gaussian-baseline students (Lipschitz 0.203 vs. 0.257; gradient norm 3.20 vs. 3.87×10^{-5}) but recover substantially less return (22.5% vs. 36.6%). The following tables disaggregate the same comparison across environments and synthetic distributions. All values are mean $\pm$ SEM across 10 seeds for full-width 100k-sample students. Lipschitz constants are mean operator norms of the policy Jacobian on real rollout observations; gradient norms are reported in units of 10^{-5}.

The shifted distribution (Gaussian with $\mu = 1$) consistently produces students with the lowest Lipschitz constants and gradient norms in every environment, yet also the lowest recovery percentages. This confirms the main-text observation that smoothness alone is insufficient: a student can be smooth and still fail if its smoothness is not aligned with the teacher's sensitive directions.

D.7 Per-environment sensitivity spectrum

Table 13 in Appendix D.3 reports detailed real-versus-synthetic sensitivity structure for Humanoid-v5. Here we report the corresponding teacher-side spectrum across environments, computed from the eigenvalue-based participation ratio of $G_\pi = (1/n)\sum_i J_i^\top J_i$ on real teacher-rollout observations. HalfCheetah-v5 is included as a fully replicated medium-dimensional locomotion environment.

D.8 Teacher–student sensitivity alignment

We test whether students recover the teacher's sensitivity structure rather than an unrelated low-dimensional approximation. For each policy, we compute coordinate-wise sensitivity

$$I_j(\pi) = \frac{1}{n}\sum_i \|J_i[:, j]\|_2^2,$$

and compare teacher and student using Spearman correlation, cosine similarity, and top-k eigenspace overlap on real teacher-rollout observations.

Alignment is already high at 25k samples and tightens with scale to Spearman correlation near 0.995 and subspace overlap $\cos\bar{\theta} \geq 0.985$. This supports the interpretation that successful students recover the teacher's dominant sensitivity structure.

Table 23 extends the coordinate-level comparison across environments.

Coordinate alignment is high across environments, showing that the Humanoid result is not specific to the highest-dimensional task. Table 24 further shows that HalfCheetah-v5 alignment tightens

Table 18: Subspace intervention recovery at 20 vs. 100 training epochs. Recovery (%), mean ± std across 10 seeds. All conditions use 10^5 samples. 100-epoch runs are independent stochastic training runs, not continuations of the 20-epoch students.

	Humanoid-v5		Ant-v5	
Condition	20 ep.	100 ep.	20 ep.	100 ep.
Full	39.8 ± 18.0	32.9 ± 16.8	81.0 ± 11.2	91.3 ± 11.6
Top-d_{sens}	16.4 ± 9.9	14.4 ± 5.0	25.8 ± 13.1	26.3 ± 9.3
Random-d_{sens}	3.6 ± 1.8	3.3 ± 1.5	18.7 ± 2.5	17.2 ± 3.4
Bottom-d_{sens}	2.8 ± 0.8	3.2 ± 1.4	16.9 ± 4.3	18.0 ± 2.3
Top-r_{90}	25.0 ± 9.1	14.9 ± 5.8	39.6 ± 19.3	28.2 ± 15.8
Random-r_{90}	4.5 ± 1.3	3.9 ± 1.2	16.5 ± 3.7	9.9 ± 3.5

Table 19: Hopper-v5 per-distribution smoothness.

Distribution	Recovery (%)	Lipschitz	$\|\text{logit}\|$	Grad. norm ($\times 10^{-5}$)
Teacher	100.0	0.696	1.516	17.07
Gaussian, $\sigma = 1$	65.3 ± 6.9	0.648 ± 0.018	1.476 ± 0.034	15.98 ± 0.38
Gaussian, $\sigma = 2$	60.9 ± 5.5	0.605 ± 0.016	1.393 ± 0.031	15.07 ± 0.35
Gaussian, $\sigma = 0.5$	53.5 ± 6.2	0.646 ± 0.019	1.537 ± 0.045	15.87 ± 0.39
Gaussian, $\mu = 1, \sigma = 1$	42.1 ± 6.2	0.598 ± 0.024	1.325 ± 0.046	14.57 ± 0.60
Uniform $[-1, 1]$	55.2 ± 6.4	0.650 ± 0.019	1.526 ± 0.038	16.03 ± 0.39
Uniform $[-0.5, 0.5]$	34.0 ± 4.1	0.618 ± 0.018	1.589 ± 0.052	15.38 ± 0.32
Laplace, $b = 1$	64.0 ± 6.7	0.637 ± 0.017	1.459 ± 0.032	15.74 ± 0.39

Table 20: Ant-v5 per-distribution smoothness.

Distribution	Recovery (%)	Lipschitz	$\|\text{logit}\|$	Grad. norm ($\times 10^{-5}$)
Teacher	100.0	0.187	0.941	2.49
Gaussian, $\sigma = 1$	56.6 ± 3.9	0.159 ± 0.005	0.889 ± 0.025	2.03 ± 0.04
Gaussian, $\sigma = 2$	53.6 ± 3.3	0.156 ± 0.007	0.865 ± 0.034	1.97 ± 0.07
Gaussian, $\sigma = 0.5$	48.4 ± 3.8	0.172 ± 0.003	0.955 ± 0.023	2.20 ± 0.04
Gaussian, $\mu = 1, \sigma = 1$	33.8 ± 0.9	0.148 ± 0.002	0.811 ± 0.004	1.91 ± 0.04
Uniform $[-1, 1]$	49.9 ± 4.2	0.171 ± 0.003	0.947 ± 0.024	2.20 ± 0.04
Uniform $[-0.5, 0.5]$	35.3 ± 3.1	0.183 ± 0.003	1.032 ± 0.027	2.36 ± 0.05
Laplace, $b = 1$	57.8 ± 3.7	0.154 ± 0.006	0.865 ± 0.028	1.96 ± 0.05

Table 21: Humanoid-v5 per-distribution smoothness.

Distribution	Recovery (%)	Lipschitz	$\|\text{logit}\|$	Grad. norm ($\times 10^{-5}$)
Teacher	100.0	0.255	3.001	4.10
Gaussian, $\sigma = 1$	36.6 ± 0.5	0.257 ± 0.004	3.274 ± 0.036	3.87 ± 0.08
Gaussian, $\sigma = 2$	32.3 ± 0.9	0.261 ± 0.008	3.302 ± 0.081	4.06 ± 0.13
Gaussian, $\sigma = 0.5$	39.5 ± 0.8	0.268 ± 0.003	3.393 ± 0.047	3.89 ± 0.07
Gaussian, $\mu = 1, \sigma = 1$	22.5 ± 1.1	0.203 ± 0.003	2.553 ± 0.028	3.20 ± 0.06
Uniform $[-1, 1]$	39.2 ± 0.8	0.265 ± 0.002	3.355 ± 0.032	3.87 ± 0.06
Uniform $[-0.5, 0.5]$	40.0 ± 0.6	0.271 ± 0.006	3.412 ± 0.087	3.86 ± 0.10
Laplace, $b = 1$	34.5 ± 0.9	0.257 ± 0.006	3.267 ± 0.065	3.95 ± 0.11

substantially at 500k samples, where Gaussian and Laplace students recover the teacher's coordinate and subspace sensitivity structure almost exactly.

D.9 Real-fit Gaussian: decomposition diagnostic

As a diagnostic baseline, we fit a multivariate Gaussian to teacher rollout observations and distill from samples drawn from this distribution. This baseline is real-data-informed and is therefore *not* part of our data-free synthetic recovery claim. We include it to characterize how much of the gap

Table 22: Teacher–student sensitivity alignment on Humanoid-v5 ($k = \lceil d_{\text{sens}} \rceil = 16$).

Distribution	Samples	Spearman ρ	Cosine sim.	Subspace $\cos\bar{\theta}$
Gaussian shifted	25k	0.939	0.991	0.947
Gaussian baseline	25k	0.981	0.996	0.986
Gaussian narrow/wide	25k	0.978–0.983	0.995–0.997	0.984–0.987
All distributions	500k	0.989–0.992	0.996–0.998	≥ 0.985
All distributions	2M	0.993–0.995	0.998–0.999	≥ 0.985

Table 23: Per-environment teacher–student coordinate-sensitivity alignment. Values are mean $\pm$ std or SEM as reported by each analysis.

Environment	Spearman ρ	Cosine sim.
Hopper-v5	0.833 ± 0.024	0.981 ± 0.002
Ant-v5	0.961 ± 0.003	0.996 ± 0.000
Humanoid-v5	0.830 ± 0.008	0.996 ± 0.000

between the data-free recovery percentages of Section 4 and full teacher performance is attributable to distributional structure available only through rollouts.

Setup. For each environment we collect $n = 10{,}000$ real teacher-rollout observations and compute their empirical mean $\hat{\mu}$ and Ledoit-Wolf-shrunk covariance $\hat{\Sigma}$ in normalized observation space. Ledoit-Wolf shrinkage (Ledoit & Wolf, 2004) regularizes the covariance estimate toward a scaled identity; for Humanoid ($d = 348$, $n = 10{,}000$) the analytical shrinkage coefficient is $\alpha \approx 0.03$, indicating the sample covariance is well-conditioned at this sample size. We compare four sampling distributions in a 2×2 factorial design:

- $\mathcal{N}(0, I)$ — the simple Gaussian baseline used throughout the paper (neither corrected center nor empirical covariance);
- $\mathcal{N}(\hat{\mu}, I)$ — centering corrected to empirical rollout mean, isotropic covariance;
- $\mathcal{N}(0, \hat{\Sigma})$ — empirical covariance, zero-centered;
- $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ — the real-fit Gaussian.

The factorial design isolates the contribution of distributional centering and empirical covariance structure, and tests whether their combination is additive. The four recovery conditions are summarized in Table 3 of the main text, while Table **??** reports the corresponding centering, covariance, and interaction decomposition.

Why centering matters at all. The simple Gaussian $\mathcal{N}(0, I)$ in normalized space is centered at the running-mean normalization statistics $(\mu_{\text{rms}}, \sigma_{\text{rms}})$, which are accumulated as running averages over the entire training history. By design, these statistics lag the converged policy's current rollout statistics: a `VecNormalize` wrapper that tracked the current rollout mean tightly would induce non-stationary value-function targets and destabilize PPO training. The empirical mean of normalized rollout observations from a converged Humanoid policy has $\|\hat{\mu}\| \approx 4.9$ in 348-dimensional normalized space, reflecting genuine drift between training-history-averaged and converged-policy state distributions. This drift is a feature of the standard normalization design, not a bug. The $2{\times}2$ design quantifies how much it costs in recovery.

Interpretation. On both environments, off-diagonal covariance structure dominates the recovery improvement: 66% of the total gap on Humanoid and 89% on Ant. Re-centering on the empirical rollout mean contributes a smaller fraction (13% Humanoid, 24% Ant). The two effects are super-additive on Humanoid (interaction $+11.4$pp) and approximately independent on Ant (interaction -2.8pp, within noise). The Humanoid super-additivity is consistent with a high-dimensional anisotropic policy: the policy-relevant region is at a particular off-diagonal direction *from* a particular center, and getting both factors right captures more than the sum of their individual contributions.

The decomposition has direct relevance to the data-free claim. The `obs_rms` normalization captures per-dimension scale but not inter-coordinate correlations; the convention of using $\mathcal{N}(0, I)$ in normal-

Table 24: HalfCheetah-v5 teacher–student sensitivity alignment. Coordinate metrics compare teacher and student coordinate-wise sensitivity vectors, while subspace overlap measures alignment of the top-$\lceil d_{\text{sens}} \rceil$ eigenspaces. Values are mean $\pm$ std across 10 seeds.

Student	Samples	Spearman ρ	Cosine sim.	Subspace $\cos\bar{\theta}$
Gaussian baseline	100k	0.84 ± 0.19	0.98 ± 0.03	0.95 ± 0.03
Gaussian baseline	500k	0.98 ± 0.02	1.00 ± 0.00	0.99 ± 0.02
Laplace baseline	500k	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00
Uniform wide	500k	0.90 ± 0.12	0.99 ± 0.01	0.97 ± 0.01

Table 25: Decomposition of recovery improvement from $\mathcal{N}(0, I)$ to $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$. Centering effect is $\mathcal{N}(\hat{\mu}, I) - \mathcal{N}(0, I)$; covariance effect is $\mathcal{N}(0, \hat{\Sigma}) - \mathcal{N}(0, I)$; interaction is the residual. Shares are percentages of the total gap.

Environment	Centering (pp)	Covariance (pp)	Interaction (pp)	Total (pp)
Humanoid-v5	$+6.7$ (13%)	$+34.8$ (66%)	$+11.4$ (21%)	$+52.8$
Ant-v5	$+5.0$ (24%)	$+18.9$ (89%)	-2.8 (-13%)	$+21.1$

ized space therefore samples from a distribution that is correctly scaled per-dimension but does not respect off-diagonal state structure. The real-fit experiment quantifies the cost of this restriction. On Humanoid, the cost is roughly 50 percentage points of recovery; the bulk of this cost (35pp) comes from not capturing the empirical covariance, which is unavailable from policy metadata alone.

D.10 Student sensitivity on synthetic vs. real inputs

Tables 19, 20, 21 and 23 measure student geometry on real teacher-rollout observations. Because students are trained exclusively on synthetic inputs, it is also informative to ask whether the student's input-sensitivity geometry differs between the distribution it was trained on and the distribution it is evaluated against.

We use a coordinate-level participation-ratio measure

$$d_{\text{coord}} = \frac{\left(\sum_j I_j\right)^2}{\sum_j I_j^2}, \qquad I_j(\pi) = \frac{1}{n} \sum_i \|J_i[:, j]\|_1,$$

which captures how broadly the student's per-coordinate sensitivity is distributed. This is a coordinate-level quantity and is distinct from the eigenvalue-based d_{sens} used in Appendix D.7; the two are not directly comparable.

Table 26: Hopper-v5 student d_{coord} on synthetic vs. real inputs. Mean $\pm$ std across 10 seeds.

Distribution	d_{coord} (synth.)	d_{coord} (real)	Δ (synth. $-$ real)
Gaussian, $\sigma = 1$	9.8 ± 0.5	9.4 ± 0.7	$+0.3$
Gaussian, $\mu = 1, \sigma = 1$	9.8 ± 0.8	9.3 ± 0.7	$+0.5$
Uniform $[-1, 1]$	9.5 ± 0.5	9.4 ± 0.7	$+0.1$

Table 27: Ant-v5 student d_{coord} on synthetic vs. real inputs. Mean $\pm$ std across 10 seeds.

Distribution	d_{coord} (synth.)	d_{coord} (real)	Δ (synth. $-$ real)
Gaussian, $\sigma = 1$	80.9 ± 2.6	76.8 ± 4.2	$+4.2$
Gaussian, $\mu = 1, \sigma = 1$	84.4 ± 2.8	78.1 ± 3.6	$+6.3$
Uniform $[-1, 1]$	79.0 ± 2.7	77.3 ± 4.5	$+1.7$

The sign of $\Delta = d_{\text{coord}}^{\text{synth}} - d_{\text{coord}}^{\text{real}}$ depends on observation dimensionality. In Hopper-v5 ($d_{\text{obs}} = 11$), Δ is small and positive: student sensitivity is slightly more diffuse on synthetic inputs, but the

Table 28: Humanoid-v5 student d_{coord} on synthetic vs. real inputs. Mean $\pm$ std across 10 seeds.

Distribution	d_{coord} (synth.)	d_{coord} (real)	Δ (synth. $-$ real)
Gaussian, $\sigma = 1$	310.6 ± 2.7	338.5 ± 1.5	-27.9
Gaussian, $\mu = 1, \sigma = 1$	315.7 ± 2.5	338.8 ± 1.5	-23.2
Uniform $[-1, 1]$	303.2 ± 3.1	336.0 ± 1.7	-32.8

magnitude is below seed-level noise. In Ant-v5 ($d_{\text{obs}} = 105$), Δ is positive and modest ($+1.7$ to $+6.3$), consistent with iid synthetic samples covering more input coordinates than correlated trajectory observations. In Humanoid-v5 ($d_{\text{obs}} = 348$), the pattern reverses and Δ is strongly negative: student sensitivity is substantially more concentrated on synthetic inputs ($\Delta \approx -23$ to -33). This dimension-dependent reversal is a direct prediction of concentration of measure in high dimensions, where iid samples in the ambient cube project tightly onto any fixed eigendirection of G_π. Appendix D.12 examines this mechanism directly.

D.11 Trajectory-control statistical analysis

This section reports the per-budget paired comparison between Variant A (contiguous trajectory chunks) and Variant B (iid samples from a buffer of teacher rollouts) used in Section 6 of the main text. The paired difference $\Delta = B - A$ is computed for each (environment, budget) pair across 10 matched student seeds. Both a paired two-sided t-test and a binomial sign test are reported.

Table 29: Hopper-v5 trajectory-control comparison. Mean $\pm$ SEM, 10 seeds per condition.

Budget (k)	Variant A (%)	Variant B (%)	Δ	$n(B > A)$	p_t	p_{sign}
0	94.1 ± 14.4	94.9 ± 13.8	$+0.8 \pm 3.0$	$8/10$	0.811	0.109
5	96.8 ± 12.5	97.5 ± 14.4	$+0.7 \pm 4.1$	$2/10$	0.874	0.109
10	95.8 ± 13.0	99.2 ± 14.2	$+3.4 \pm 2.9$	$5/10$	0.299	1.000
20	96.9 ± 14.6	97.4 ± 14.8	$+0.5 \pm 1.9$	$5/10$	0.791	1.000
50	97.1 ± 14.2	98.3 ± 13.9	$+1.2 \pm 1.8$	$4/10$	0.542	0.754
100	100.0 ± 14.1	99.4 ± 14.5	-0.6 ± 1.4	$4/10$	0.694	0.754

Table 30: HalfCheetah-v5 trajectory-control comparison. Mean $\pm$ SEM, 10 seeds per condition.

Budget (k)	Variant A (%)	Variant B (%)	Δ	$n(B > A)$	p_t	p_{sign}
0	81.0 ± 11.9	81.0 ± 11.9	$+0.0 \pm 0.0$	–	–	–
5	93.6 ± 11.2	90.6 ± 12.1	-3.0 ± 5.6	$5/10$	0.609	1.000
10	93.7 ± 9.9	94.4 ± 10.9	$+0.8 \pm 5.0$	$4/10$	0.878	0.754
20	95.7 ± 12.5	91.5 ± 13.6	-4.1 ± 7.4	$4/10$	0.592	0.754
50	93.2 ± 15.0	96.5 ± 12.8	$+3.3 \pm 6.8$	$5/10$	0.639	1.000
100	89.9 ± 11.0	92.6 ± 11.3	$+2.7 \pm 3.8$	$4/10$	0.503	0.754

Table 31: Ant-v5 trajectory-control comparison. Mean $\pm$ SEM, 10 seeds per condition.

Budget (k)	Variant A (%)	Variant B (%)	Δ	$n(B > A)$	p_t	p_{sign}
0	88.9 ± 2.7	88.1 ± 3.0	-0.8 ± 2.6	$5/10$	0.774	1.000
5	97.8 ± 2.2	98.8 ± 2.2	$+1.0 \pm 1.3$	$6/10$	0.471	0.754
10	98.5 ± 2.0	98.9 ± 2.2	$+0.4 \pm 1.6$	$4/10$	0.831	0.754
20	98.7 ± 2.1	99.3 ± 2.4	$+0.5 \pm 1.8$	$6/10$	0.788	0.754
50	100.4 ± 1.9	96.4 ± 2.5	-4.0 ± 1.8	$3/10$	0.064	0.344
100	100.0 ± 1.9	99.4 ± 2.3	-0.6 ± 1.8	$5/10$	0.758	1.000

No (environment, budget) combination reaches significance at $\alpha = 0.05$ on either test. Sign-test counts cluster around $5/10$ across all conditions, consistent with the null hypothesis of no systematic preference. With $n = 10$ seeds and an empirical residual standard deviation of approximately $\sigma \approx 4$ percentage points, the paired test has $\sim 80\%$ power to detect differences of 2–3 percentage points

Table 32: Humanoid-v5 trajectory-control comparison. Mean ± SEM, 10 seeds per condition.

Budget (k)	Variant A (%)	Variant B (%)	Δ	$n(B > A)$	p_t	p_{sign}
0	35.4 ± 6.5	37.4 ± 6.7	$+2.0 \pm 1.1$	$8/10$	0.124	0.109
5	80.1 ± 5.8	81.4 ± 6.6	$+1.3 \pm 2.9$	$6/10$	0.685	0.754
10	86.3 ± 5.4	91.1 ± 3.9	$+4.8 \pm 4.1$	$5/10$	0.287	1.000
20	93.6 ± 3.8	92.1 ± 4.5	-1.5 ± 2.3	$4/10$	0.543	0.754
50	98.5 ± 3.4	95.4 ± 4.0	-3.0 ± 1.8	$5/10$	0.142	1.000
100	100.0 ± 3.8	102.4 ± 3.0	$+2.4 \pm 2.4$	$5/10$	0.367	1.000

(two-sided $\alpha = 0.05$). The observed effect sizes lie below this detection threshold, indicating that a substantially larger study would be needed to identify any small residual trajectory-structure effect.

D.12 Projection analysis for uniform-wide sampling

A consistent finding in our experiments is that students trained on Uniform $[-1, 1]$ synthetic observations achieve recovery rates competitive with Gaussian and Laplace baselines, even though uniform-wide samples are far from the empirical state distribution in ambient distance. This subsection examines why large ambient distance need not imply large displacement along a fixed sensitive direction.

Setup. We compute the teacher Gram matrix $G_\pi = \mathbb{E}_{x\sim\text{real}}[J(x)^\top J(x)] \in \mathbb{R}^{348\times348}$ for the Humanoid-v5 teacher, where $J(x)$ is the Jacobian of the teacher's deterministic action mean. G_π is estimated from 120 real teacher-rollout observations using finite differences ($\epsilon = 10^{-4}$). We project two sets of observations onto the top four eigenvectors of G_π: 400 real teacher-rollout observations, and 800 samples from the synthetic distribution Uniform $[-1, 1]^{348}$. The top four eigenvectors capture 31.5% of the total Jacobian variance and are chosen for visual clarity; the full sensitive subspace, as reported in Appendix D.7, has $d_{\text{sens}} = 15.5$ and the top 78 eigendirections capture 90% of the variance.

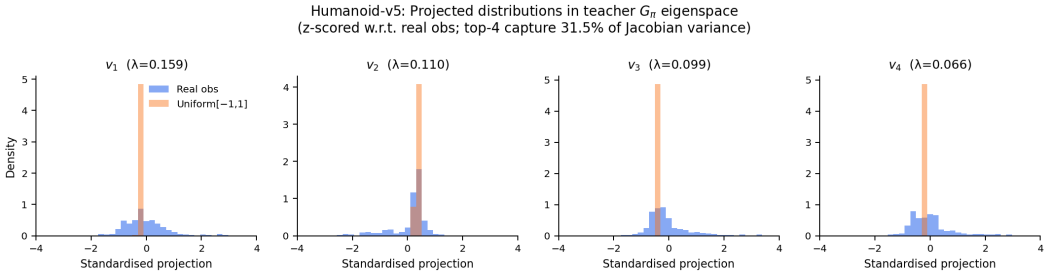

Figure 4: Projections of real (blue) and uniform-wide synthetic (orange) observations onto the top four eigenvectors of the Humanoid-v5 teacher Gram matrix. Uniform samples have bounded, order-one projections despite their large ambient norm.

Table 33: Mean and standard deviation of the projection of real and synthetic Humanoid observations onto the top four eigenvectors of G_π.

	λ (Jacobian var.)	Real (mean ± std)	Synthetic (mean ± std)
v_1	0.159	$+9.52 \pm 35.71$	-0.002 ± 0.585
v_2	0.110	-16.45 ± 51.42	-0.015 ± 0.577
v_3	0.099	$+12.94 \pm 28.53$	$+0.001 \pm 0.585$
v_4	0.066	$+9.81 \pm 44.15$	$+0.026 \pm 0.570$

Discussion. Let $x \in [-1, 1]^d$ have independent coordinates and let v be any fixed unit vector. Then

$$\mathbb{E}[v^\top x] = 0, \qquad \text{Var}(v^\top x) = \frac{1}{3}\sum_{j=1}^{d} v_j^2 = \frac{1}{3}.$$

Thus the projection standard deviation is $1/\sqrt{3} \approx 0.577$, independent of d, exactly as observed in Table 33. Meanwhile, the expected squared ambient norm is $d/3$. Uniform samples can therefore be far from rollout states in Euclidean distance while maintaining bounded, nonzero variation along every fixed teacher-sensitive direction.

This calculation does not show that uniform samples approximate the rollout distribution; Table 33 shows that they do not. Instead, it clarifies why ambient distance is a poor measure of useful query coverage. Uniform sampling excites every sensitive direction at a stable scale and also provides full ambient support. These are the two properties isolated by the subspace and thickening interventions.

In our experiments, centered distributions with sufficient variation along teacher-sensitive directions and nonzero ambient support produce similar recovery rates. Gaussian $\sigma = 1$, Laplace $b = 1$, and Uniform $[-1, 1]$ all satisfy these conditions for the environments studied (Appendix D.6). This is consistent with the dimension-dependent reversal of Δ in Appendix D.10: in high dimensions, synthetic inputs project onto a smaller effective subspace than real ones, but this does not impede recovery because the projection still covers the policy-relevant directions.

E Breadth Checks Beyond PPO Locomotion

The experiments in this section test whether the two geometric interventions depend on the main paper's combination of PPO, tanh actors, continuous actions, and MuJoCo locomotion.

E.1 Discrete-action ReLU policy

We train 10 PPO teachers on LunarLander with an eight-dimensional input, discrete actions, and a ReLU MLP. The sensitivity dimension is $d_{\text{sens}} = 1.52$, with $r_{90} = r_{95} = 2$. At 10k passive queries, Gaussian, uniform, and Laplace students recover respectively $105.0 \pm 5.4\%$, $102.3 \pm 6.7\%$, and $106.9 \pm 6.6\%$ of teacher return.

Table 34: LunarLander subspace and thickening interventions.

Condition	Normalized overlap	Recovery (%)
Full Gaussian	1.000	106.2 ± 2.2
Top-2	0.969	54.8 ± 16.1
Random-2	0.315	-102.1 ± 51.0
Bottom-2	0.004	-329.9 ± 100.8
Top-2, $\varepsilon = 0$	0.969	52.0 ± 11.6
Top-2, $\varepsilon = 0.05$	0.969	70.4 ± 14.8
Top-2, $\varepsilon = 0.10$	0.969	89.9 ± 10.0
Top-2, $\varepsilon = 0.25$	0.971	100.1 ± 4.6
Top-2, $\varepsilon = 0.50$	0.977	103.8 ± 5.1
Top-2, $\varepsilon = 1.00$	1.000	107.9 ± 4.0

Top-2 is substantially better than matched random and bottom subspaces. Orthogonal thickening raises recovery while leaving directional overlap nearly constant. Thus both factors replicate with a discrete action representation and without tanh activations.

E.2 Off-policy SAC actor

We repeat the intervention suite using 10 stochastic SAC teachers on Ant-v5 and distill its deterministic action mean. The teachers have input dimension 105, $d_{\text{sens}} = 2.87$, $r_{90} = 7$, and $r_{95} = 16$. Passive recovery at 500k queries is $118.9 \pm 2.3\%$ for Gaussian, $103.2 \pm 8.5\%$ for uniform, and $69.7 \pm 17.1\%$ for Laplace inputs.

The strict subspace ordering and thickening response match the PPO results. Smoothness alone again fails as an explanation: Gaussian, uniform, and Laplace students have Lipschitz estimates 3.009, 3.842, and 2.486, but recover 69.3%, 96.6%, and 22.3%, respectively.

Table 35: SAC Ant intervention. Held-out MSE is measured on teacher-rollout observations.

Condition	Overlap	Recovery (%)	Action MSE
Top-3	0.818	-6.8 ± 2.1	–
Random-3	0.028	-20.1 ± 17.4	–
Bottom-3	0.000	-63.7 ± 36.1	–
Top-7, $\varepsilon = 0$	0.903	-6.0 ± 1.7	1.313
Top-7, $\varepsilon = 0.05$	0.903	35.6 ± 13.1	0.303
Top-7, $\varepsilon = 0.10$	0.904	71.2 ± 11.7	0.212
Top-7, $\varepsilon = 0.25$	0.909	63.5 ± 5.9	0.154
Top-7, $\varepsilon = 0.50$	0.927	81.1 ± 17.0	0.119
Top-7, $\varepsilon = 1.00$	1.000	77.9 ± 6.6	0.115

E.3 MNIST COMPLEXITY BOUNDARY

To separate ambient dimension from functional complexity, we train fully connected ReLU teachers on MNIST while shuffling increasing fractions of the training labels. Each level uses 20 independently trained teachers. Training accuracy remains near 98% through 75% shuffling, but effective sensitivity rank rises and passive synthetic recovery falls.

Table 36: MNIST label-shuffling experiment. Match is student-teacher class agreement. Values are mean $\pm$ standard deviation over 20 teachers.

Shuffle	Teacher train	Teacher test	d_{sens}	Synthetic match
0%	98.36 ± 0.22	96.92 ± 0.19	6.10 ± 0.32	89.98 ± 1.96
10%	98.25 ± 0.14	92.22 ± 0.49	19.68 ± 0.71	44.47 ± 5.79
25%	98.26 ± 0.20	81.71 ± 0.95	29.10 ± 1.70	34.02 ± 3.54
50%	98.23 ± 0.21	59.78 ± 1.30	42.32 ± 2.11	24.05 ± 2.90
75%	98.23 ± 0.18	34.57 ± 0.83	56.61 ± 2.85	14.95 ± 1.33
100%	96.27 ± 2.17	9.88 ± 0.37	67.12 ± 3.78	10.71 ± 0.17

MNIST has 784 input dimensions, yet the clean teacher reaches about 90% synthetic match because its effective rank is low. Conversely, only 10% label shuffling increases effective rank from 6.10 to 19.68 and reduces synthetic match from 89.98% to 44.47%. Ambient dimension alone therefore does not determine recoverability. This experiment supports the role of effective functional complexity, but it does not test closed-loop behavior.

F CAPACITY AND QUERY-BUDGET SCALING

This appendix asks whether action imitation error can be predicted from student parameter count P, synthetic-query budget N, and teacher sensitivity dimension d_{sens}. It does not claim a scaling law for episodic return.

F.1 UNIFIED ANT STUDY

The controlled Ant-v5 study uses ten independently trained PPO teachers, student widths 64, 128, and 256, and budgets from 10k to 2M. Each teacher-condition pair has one independently initialized student, and each teacher and student is evaluated for 100 paired episodes. Budgets are nested prefixes of one synthetic dataset per teacher, and student initialization is paired across budgets. This gives 240 student conditions.

On held-out normalized teacher-rollout observations, the Ant-only fit is

$$\begin{aligned}\log(\text{MSE}) = &-4.6158 - 0.3577 \log_2(P/100{,}000) - 0.2537 \log_2(N/100{,}000) \\ &+ 0.3540 \log_2(d_{\text{sens}}/10),\end{aligned} \tag{7}$$

with $R^2 = 0.840$. Per doubling, the fitted multipliers are 0.699 for P, 0.776 for N, and 1.425 for d_{sens}. The sensitivity term adds only modest information: removing it reduces R^2 from 0.8404 to

0.8362 and raises leave-one-teacher-out squared log-error from 0.1174 to 0.1194. A teacher-clustered bootstrap gives its coefficient a 95% interval of approximately $[0.003, 0.561]$ on the natural-log-MSE scale per doubling.

Fitting only through 500k and testing at 1M gives median multiplicative errors of $1.77\times$, $1.50\times$, and $1.69\times$ for widths 64, 128, and 256. At 2M these errors rise to $2.34\times$, $1.85\times$, and $2.28\times$. Equation 7 is therefore an empirical within-environment approximation, not an asymptotic sample-complexity law.

F.2 Cross-environment check

We compare Ant, Hopper, and HalfCheetah over their common 10k, 25k, 50k, and 100k budgets, giving 120 conditions and ten teacher clusters per environment. The dependent variable remains clipped action MSE. Environment-specific intercepts absorb differences in observation dimension, action dimension, and action-error scale; these dimensional effects are not independently estimated.

Table 37: Cross-environment models on 360 student conditions. LOOTO reports leave-one-teacher-out squared log-error.

Model	R^2	LOOTO error
Shared P, N slopes	0.782	0.231
Shared P, N, d_{sens} slopes	0.796	0.225
Environment-specific slopes	0.903	0.120

The total environment-specific coefficients are:

Environment	$\log_2 P$	$\log_2 N$	$\log_2 d_{\text{sens}}$
Ant	-0.164	-0.325	$+0.332$
Hopper	-0.457	-0.830	$+1.575$
HalfCheetah	-0.408	-0.579	$+0.152$

Adding sensitivity to the shared-slope model increases R^2 by 0.014. The current 30-teacher clustered interval for the shared sensitivity coefficient is $[0.006, 1.339]$, indicating a positive estimate with poorly constrained magnitude. Allowing environment-specific slopes gives the much larger improvement. Calibrating only a target-environment intercept, the Ant coefficients predict Hopper with median multiplicative error $2.06\times$ and HalfCheetah with error $1.39\times$. The qualitative effects transfer, but the numerical slopes are not universal. The modest predictive contribution of d_{sens} does not contradict the directional interventions in Section 5. Here, d_{sens} is a scalar summary of teacher complexity and is used to predict action-error scaling across teachers and budgets. In contrast, the subspace experiments directly manipulate which teacher-sensitive directions are covered while controlling support dimension and per-direction variance. Thus, d_{sens} need not be a strong scalar predictor for directional coverage itself to have a causal effect on recovery.

F.3 Action error and closed-loop return separate

Hopper provides the clearest saturation case. From 250k to 500k queries, mean action MSE falls from 0.0105 to 0.0051 at width 16, from 0.00214 to 0.00173 at width 32, and from 0.000653 to 0.000350 at width 64, while mean return recovery remains near 98–100%. In HalfCheetah, increasing the common-range budget from 10k to 100k reduces MSE at every width while recovery changes sharply from negative values toward teacher performance.

These results support two limited claims. Larger synthetic budgets and student capacity consistently reduce action error, and teacher sensitivity contributes some information about imitation difficulty. They do not support a universal law for return or for the sample count needed to reach a fixed recovery level. Closed-loop dynamics, action-error direction, and return saturation remain environment-dependent.

G Coverage Guarantees and Synthetic-Pilot Experiments

This section gives complete proofs for the coverage statements in Section 5.1 and full results for the synthetic-only query-design experiment in Section 5.5. For clarity, let q be the query distribution, let ρ be the deployment distribution, and define their second-moment matrices $\Sigma_q = \mathbb{E}_{x\sim q}[xx^\top]$

and $\Sigma_\rho = \mathbb{E}_{x\sim\rho}[xx^\top]$. For affine policies, the statements apply after augmenting x with a constant coordinate.

G.1 Strict-subspace non-identifiability

Proposition 1. Suppose that q is supported on a proper linear subspace $S \subsetneq \mathbb{R}^d$. If there is a vector $v \in S^\perp$ such that $v^\top \Sigma_\rho v > 0$, then two policies can agree on every query while having arbitrarily different deployment error.

Proof. Fix a nonzero output vector u and define $f(x) = 0$ and $g_c(x) = cuv^\top x$ for any $c > 0$. For every $x \in S$, $v^\top x = 0$, so $f(x) = g_c(x)$ and their query loss is zero. In contrast,

$$\mathbb{E}_\rho \|f(x) - g_c(x)\|_2^2 = c^2 \|u\|_2^2 v^\top \Sigma_\rho v > 0.$$

This error grows without bound as c increases. □

G.2 Deterministic covariance-transfer bound

Proposition 2. Let the local student-teacher error be $e(x) = Ax$, where $A \in \mathbb{R}^{m\times d}$, and suppose $\Sigma_q \succ 0$. Define

$$L_q(A) = \mathrm{Tr}(A\Sigma_q A^\top), \qquad L_\rho(A) = \mathrm{Tr}(A\Sigma_\rho A^\top).$$

Then

$$L_\rho(A) \le \lambda_{\max}\left(\Sigma_q^{-1/2}\Sigma_\rho\Sigma_q^{-1/2}\right) L_q(A). \tag{8}$$

If Σ_q is singular and its null space contains a direction with positive deployment variance, no finite bound of this form exists.

Proof. Let $M = \Sigma_q^{-1/2}\Sigma_\rho\Sigma_q^{-1/2}$ and $\gamma = \lambda_{\max}(M)$. Since $M \preceq \gamma I$, congruence with $\Sigma_q^{1/2}$ gives $\Sigma_\rho \preceq \gamma\Sigma_q$. Congruence with A and monotonicity of trace give $L_\rho(A) \le \gamma L_q(A)$. The singular case follows from Proposition 1 by taking $e(x) = cuv^\top x$ in an uncovered direction. □

This bound is deterministic. It does not assume noisy teacher labels. It also shows why nonzero support alone can be inadequate: the factor in Eq. 8 can be very large when query variance is small along a deployment-relevant direction.

G.3 FINITE-SAMPLE INVERSE COVERAGE

Proposition 3. Consider the local linear model $y = Wx + \xi$, with $x_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_q)$, $\Sigma_q \succ 0$, and $\xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_m)$ independent of x_i. Let $\widehat{W}$ be ordinary least squares using $N > d + 1$ samples. Then

$$\mathbb{E}\Big[\mathrm{Tr}\Big((\widehat{W} - W)\Sigma_\rho(\widehat{W} - W)^\top\Big)\Big] = \frac{m\sigma^2}{N - d - 1}\,\mathrm{Tr}(\Sigma_\rho \Sigma_q^{-1}). \tag{9}$$

Proof. Let $X \in \mathbb{R}^{N\times d}$ be the design matrix. Conditional on X, the coefficient error for each output has covariance $\sigma^2(X^\top X)^{-1}$. Summing the deployment loss over the m outputs gives

$$\mathbb{E}[L_\rho(\widehat{W} - W) \mid X] = m\sigma^2\,\mathrm{Tr}\left(\Sigma_\rho (X^\top X)^{-1}\right).$$

Because $X^\top X$ is Wishart, $\mathbb{E}[(X^\top X)^{-1}] = \Sigma_q^{-1}/(N - d - 1)$. Substitution proves Eq. 9. □

We therefore define the inverse-coverage diagnostic

$$\mathcal{R}_{\text{inv}}(q, \rho) = \mathrm{Tr}(\Sigma_\rho \Sigma_q^{-1}). \tag{10}$$

The experimental labels are deterministic. Thus, Eq. 9 is not a literal label-noise model for our neural students. It is a local approximation in which optimization error, nonlinear approximation error, or local model misspecification acts as an effective residual. Proposition 2 is the assumption-free result for the local linear error itself.

G.4 THICKENING AND THE ISOTROPIC MINIMAX BASELINE

Let P_k project onto a k-dimensional pilot subspace and let $P_\perp = I - P_k$. For the thickened query covariance

$$\Sigma_q = s^2(P_k + \varepsilon^2 P_\perp), \qquad \varepsilon > 0,$$

Eq. 10 becomes

$$\mathcal{R}_{\text{inv}} = \frac{1}{s^2}\left[\mathrm{Tr}(\Sigma_\rho P_k) + \varepsilon^{-2}\,\mathrm{Tr}(\Sigma_\rho P_\perp)\right]. \tag{11}$$

Thus, deployment variance outside the selected subspace is amplified by $1/\varepsilon^2$. For $\Sigma_\rho = I$ and fixed query energy $\mathrm{Tr}(\Sigma_q) = d$, we have

$$\mathcal{R}_{\text{inv}} = \frac{[k + \varepsilon^2(d - k)][k + (d - k)/\varepsilon^2]}{d} \geq d, \tag{12}$$

where equality holds at $\varepsilon = 1$. The inequality follows from $t + t^{-1} \geq 2$ with $t = \varepsilon^2$. Hence aggressive concentration can diverge as $\varepsilon \to 0$.

More generally, suppose the deployment direction is unknown and can be any rank-one covariance $uu^\top$ with $\|u\|_2 = 1$. For a fixed query-energy budget $\mathrm{Tr}(\Sigma_q) = \tau$,

$$\sup_{\|u\|_2=1} u^\top \Sigma_q^{-1} u = \lambda_{\min}(\Sigma_q)^{-1} \geq d/\tau.$$

Equality holds only for $\Sigma_q = (\tau/d)I$. Isotropic querying is therefore minimax when deployment orientation is unknown. A directional design can help only when its estimate of deployment-relevant geometry is sufficiently accurate and its tail variance remains large enough.

G.5 SYNTHETIC-ONLY PILOT-GUIDED QUERY DESIGN

For each of ten existing teachers per environment, we draw 1,000 $\mathcal{N}(0, I)$ pilot inputs, compute teacher Jacobians only at these synthetic locations, select a top eigenspace, and construct a thickened subspace distribution. No rollout observation is used for design or student training. The matched action-label budget is 100,000: 1,000 pilot queries plus 99,000 distillation queries for the pilot arm, and 100,000 distillation queries for the isotropic arm. Jacobian autodifferentiation at the pilot locations is an additional disclosed white-box cost. Each saved student and its paired teacher are evaluated for 100 episodes.

The pilot-guided design does not reliably beat isotropic Gaussian querying. It is neutral on Ant and Humanoid and significantly worse on HalfCheetah and Hopper. High pilot-rollout subspace overlap is therefore not sufficient.

Table 38: Synthetic-only query design. Mean ± standard deviation across ten teachers. Rollout action MSE is a post-design diagnostic.

Environment	Query distribution	Return	Recovery (%)	Rollout MSE
Ant	Isotropic	4140.4 ± 466.0	84.0 ± 8.5	0.00974 ± 0.00121
Ant	Pilot thickened	4209.1 ± 401.9	85.3 ± 6.6	0.01235 ± 0.00110
HalfCheetah	Isotropic	1152.2 ± 181.9	68.5 ± 9.4	0.01427 ± 0.00234
HalfCheetah	Pilot thickened	726.5 ± 527.9	42.9 ± 30.8	0.04965 ± 0.01241
Hopper	Isotropic	1824.5 ± 728.2	94.0 ± 21.2	0.00771 ± 0.00306
Hopper	Pilot thickened	1239.5 ± 829.4	63.2 ± 28.5	0.10964 ± 0.05553
Humanoid	Isotropic	2178.0 ± 1008.6	33.6 ± 13.7	0.12667 ± 0.01941
Humanoid	Pilot thickened	2219.8 ± 1509.7	34.2 ± 21.2	0.10872 ± 0.01596

Table 39: Pilot geometry and paired inference. CIs use a teacher-cluster bootstrap. Rollout overlap is diagnostic and was not used for design.

Environment	k	Pilot d_{sens}	Overlap	Δ rec. (pp)	95% CI	p
Ant	43.2 ± 2.0	10.48 ± 1.15	0.806 ± 0.006	$+1.38$	$[-1.91, 4.80]$	.4629
HalfCheetah	6.4 ± 0.7	4.22 ± 0.65	0.903 ± 0.036	-25.55	$[-44.71, -8.90]$	.0059
Hopper	4.3 ± 0.7	3.59 ± 0.41	0.934 ± 0.075	-30.83	$[-48.20, -14.67]$	.0078
Humanoid	157.8 ± 2.5	17.52 ± 0.78	0.652 ± 0.012	$+0.54$	$[-4.27, 6.25]$	.8652

G.6 Pilot-size ablation

We test whether the synthetic-only pilot results are limited by the number of Jacobian-query locations. We vary the pilot budget over $\{500, 1000, 2000, 5000, 10000\}$ while keeping the concentrated query design fixed at $\eta = 0.90$ and $\varepsilon = 0.10$.

Table 40: Pilot-size ablation for the synthetic-only query design. Entries are mean paired recovery differences (percentage points) relative to isotropic querying across the same ten teachers. Increasing the pilot size does not repair the HalfCheetah or Hopper failures or improve the neutral Humanoid result.

Environment	500	1k	2k	5k	10k
HalfCheetah-v5	-27.72	-30.49	-32.61	-30.25	-45.47
Hopper-v5	-29.47	-26.21	-26.39	-27.04	-27.05
Humanoid-v5	$+4.85$	$+3.84$	$+3.70$	$+3.29$	$+0.87$

As Table 40 shows, increasing the synthetic Jacobian pilot by twenty-fold does not qualitatively change the result. All HalfCheetah conditions remain significantly worse than isotropic querying ($p = 0.0020$), while all Hopper conditions remain harmful and all but the 1k condition are significant. Humanoid remains neutral, with no monotonic benefit from increasing the pilot size. The selected rank is also stable across budgets, remaining approximately 6.4 for HalfCheetah and 4.3 for Hopper, while Humanoid stabilizes from approximately 152 to 163. Thus, the failure of the concentrated $\eta = 0.90, \varepsilon = 0.10$ design is not explained by an undersized pilot.

G.7 Tail-coverage ablation and inverse-coverage test

We vary the retained pilot energy $\eta \in \{0.90, 0.95, 0.99\}$ and orthogonal scale $\varepsilon \in \{0.10, 0.25, 0.50, 1.00\}$ on HalfCheetah and Hopper. Table 42 reports the paired recovery difference from the isotropic baseline. Increasing rank or tail variance usually repairs the failure, but it does not produce a reliable improvement over isotropic.

Finally, we compute $\mathcal{R}_{\text{inv}}$ using post-design rollout covariance and the exact analytic covariance of each candidate query distribution. This use of rollout data is diagnostic only. It does not affect rank selection, sampling, or student training.

Inverse coverage strongly predicts action MSE in Hopper and HalfCheetah, where aggressive subspace concentration fails. It is moderately predictive in Ant and uninformative in Humanoid. Raw subspace overlap does not predict the HalfCheetah or Hopper failures. Thus, $\mathcal{R}_{\text{inv}}$ is a deployment-geometry-dependent diagnostic, not a universal scalar law.

Table 41: Synthetic-only pilot design and post-design diagnostics over ten teachers per environment. Δ is paired recovery change from isotropic querying. Correlations predict rollout action MSE over the tail-coverage sweep.

Environment	Δ recovery (pp)	$\rho(\mathcal{R}_{\text{inv}}, \text{MSE})$	$\rho(\text{overlap}, \text{MSE})$
Ant	$+1.38$	0.360	0.258
HalfCheetah	-25.55	0.796	-0.108
Hopper	-30.83	0.861	-0.108
Humanoid	$+0.54$	0.002	0.334

Table 42: Synthetic-pilot tail-coverage ablation. Each cell is recovery change in percentage points relative to isotropic querying. Parentheses give rollout action MSE.

Environment	η	$\varepsilon = .10$	$.25$	$.50$	1.00
HalfCheetah	.90	-34.41 (0.0545)	-20.70 (0.0274)	-11.16 (0.0177)	-4.67 (0.0151)
HalfCheetah	.95	-24.89 (0.0354)	-1.78 (0.0210)	-1.94 (0.0157)	-0.52 (0.0133)
HalfCheetah	.99	-2.05 (0.0213)	-2.35 (0.0168)	-0.04 (0.0133)	-2.65 (0.0158)
Hopper	.90	-24.63 (0.1081)	-16.22 (0.0447)	$+2.59$ (0.0148)	-1.50 (0.0076)
Hopper	.95	-1.88 (0.0700)	-8.81 (0.0334)	-6.02 (0.0146)	-3.31 (0.0078)
Hopper	.99	-12.94 (0.0234)	-3.57 (0.0143)	-3.15 (0.0078)	$+0.82$ (0.0077)

Table 43: Prediction of held-out rollout action MSE. LOTO holds out one teacher at a time.

	Spearman with MSE		LOTO R^2	
Environment	$\mathcal{R}_{\text{inv}}$	Overlap	$\mathcal{R}_{\text{inv}}$	Overlap
Hopper	0.861	-0.108	0.526	-0.079
HalfCheetah	0.796	-0.108	0.640	-0.013
Ant	0.360	0.258	0.268	-0.046
Humanoid	0.002	0.334	-0.211	-0.143