BRIDGE: ROBUSTNESS AS DYNAMIC GENERALIZATION WITHOUT PARAMETER UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust reinforcement learning (RL) uses max-min optimization to withstand worst-case disturbances, often sacrificing nominal performance even when the encountered environment admits a less conservative policy. We introduce *BRIDGE* (Bridging Robustness and In-context Decision GEneralization), which provides an alternative perspective on robustness as dynamic generalization to infer the disturbance from context and *activate a tailored policy without parameter updates*. In-context RL (ICRL) provides a natural implementation, but requires pretraining experience that teaches how disturbances affect the environment and how to counteract them. Our key insight is that the max-min training of canonical robust RL already generates this experience through a curriculum of adversaries and specialized intermediate policies. Motivated by this, we propose the *CORE* pretraining pipeline to distill these experts into an ICRL model and use the final robust controller to collect its initial deployment context. These initial interactions preserve the robust baseline's behavior with minimal risk while providing context for tailored control. Having used ICRL to reduce conservatism around a single nominal task, we take BRIDGE one step further: using adversarial experience to strengthen ICRL's own generalization to unseen tasks under deployment disturbances. This introduces a critical challenge that the context must now reveal both what a new task requires and how its execution is disturbed. To support this joint inference, we propose the *ADAPT* architecture for policy transformers to learn to disentangle task and disturbance information from CORE curricula across training tasks. Our analysis quantifies when contextual specialization improves worst-case return over a conventional robust policy, including on unseen nominal tasks. Experiments on Dark Room, Meta-World, and 11 continuous-control tasks demonstrate improved robust control and unseen-task generalization, with and without disturbances.

1 INTRODUCTION

Reinforcement learning (RL) has demonstrated significant success in diverse domains (Barto et al., 1989), yet RL policies remain highly sensitive to environmental disturbances, such as unexpected wind dynamics or friction variations (Pinto et al., 2017). To address this, canonical *Robust RL* optimizes for worst-case performance via max-min optimization (Moos et al., 2022; Hsu & Pajic, 2024). However, when a policy must hedge across plausible perturbations, protecting worst-case performance can sacrifice nominal and average-case return (Dong et al., 2025b). The environment actually encountered may admit a much less conservative strategy.

An Alternative Perspective on Robustness. To address this challenge, we propose that robustness can instead be formulated as a *dynamic generalization problem*: we can avoid worst-case compromises if an agent can infer the hidden disturbance from its immediate context and dynamically deploy a *tailored* policy. Realizing this perspective requires adapting behavior as deployment experience reveals the environment. Yet updating model or policy parameters during deployment can be expensive and inconvenient, and may introduce instability even with regularization. We therefore seek adaptation through context *without parameter updates*, so that the agent can respond to the encountered environment while retaining its pretrained parameters. We call this perspective ***B**ridging **R**obustness and **I**n-context **D**ecision **GE**neralization* (BRIDGE).

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.

ICRL as the Natural Implementation. *In-Context Reinforcement Learning* (ICRL) provides a natural mechanism for this form of adaptation (Laskin et al., 2023; Lee et al., 2023). ICRL pretrains a transformer across a family of tasks to generate actions conditioned on a *context dataset* collected from the deployment environment. This allows a single model to express tailored behavior as its context changes, with all parameters fixed. Applying this mechanism to robust control requires experience (pretraining data) that teaches both how disturbances affect the environment and how to counteract them. The learned response must also generalize to disturbances absent from pretraining. Our key insight is that *the canonical max-min training process of robust RL already generates this experience*: a curriculum of adversaries and corresponding expert policies.

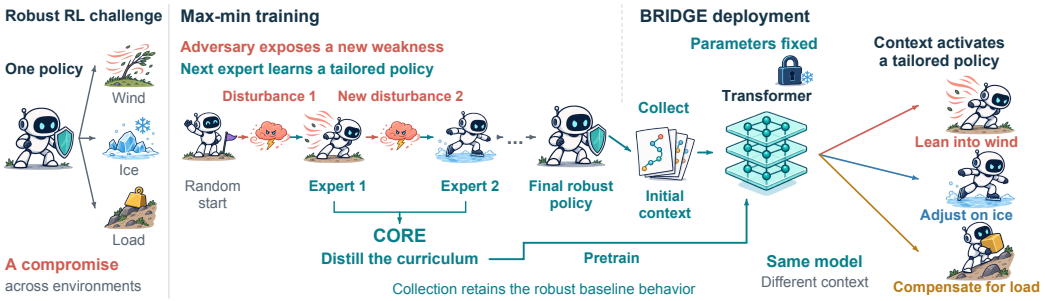


Figure 1: **BRIDGE Paradigm.** CORE distills intermediate experts; the final robust controller collects context for the frozen transformer to activate a tailored policy.

CORE: distilling the adversarial curriculum. We first consider the canonical robust RL problem around a fixed nominal task, where *nominal* denotes the task before an additional disturbance is applied. During iterative max-min optimization, each adversary exposes a weakness in the current policy, and the next expert learns to perform well under that disturbance (Pinto et al., 2017). While conventional robust RL deployment retains only the *final* robust controller, we consider the intermediate policies as specialized experts for their corresponding intermediate adversaries. With this motivation, we introduce the ***C**urriculum of **O**ptimal **R**obust **E**xperts* (CORE) pretraining pipeline. CORE combines initial task sampling with sequential adversarial search and distills the resulting expertise into an ICRL model (Section 3). Adversarial optimization thereby supplies more than an endpoint policy: it generates the environmental variation and tailored responses from which in-context generalization can be learned.

Robust context collection. Activating this expertise requires context that is not yet available when deployment begins. Collecting it with an unadapted policy can reintroduce the vulnerability that robust control is meant to address. To this end, we employ the final robust controller as the behavioral policy for initial context collection so that these interactions retain the robust baseline's behavior. The pretrained transformer then conditions on the collected experience to produce a tailored response (Section 3). BRIDGE thus uses both adversarial training's *intermediate expertise* and *final controller*.

ADAPT: robust generalization to unseen tasks. Having used ICRL to improve robust control around a single nominal task, we take BRIDGE one step further and address the robust generalization problem of ICRL itself. Specifically, after pretraining on a family of tasks, the ICRL models must generalize to *unseen tasks under deployment disturbances*. To realize this, we pretrain across tasks using their CORE curricula, then evaluate on held-out tasks absent from pretraining, the same setup as ICRL. However, during deployment, the policy must infer what the new task requires as well as how its execution is disturbed, entirely through context. This joint inference problem motivates the ***A**dversary-aware **Du**A**l**-stream **P**olicy **T**ransformer* (ADAPT). Its separate task and adversary representations encourage transferable task information to remain useful across disturbances while retaining evidence needed to tailor the response (Section 4). Adversarial variations also augment the pretraining distribution, creating an opportunity to improve unseen-task generalization even when deployment is undisturbed. Therefore, a key insight of this work is an important connection with **reciprocal reinforcement**: *ICRL can be used to address the conservatism of robust learning, while adversarial robust RL supplies experience that strengthens ICRL's own generalization.*

Performance guarantees and empirical evidence. Our analysis quantifies *when* contextual specialization improves worst-case return over a conventional robust policy, separating the available advantage from imperfect expertise, unresolved context uncertainty, and finite pretraining. It then extends the guarantee to fresh nominal tasks, connecting both settings through the same generalization principle (Section 5). Experiments on sparse-reward navigation (Laskin et al., 2023), Meta-World (Yu et al., 2019), and 11 continuous-control tasks (Todorov et al., 2012; Tunyasuvunakool et al., 2020) demonstrate improved adaptability and robustness. Curriculum, initialization, and representation ablations connect these gains to the proposed mechanisms, while evaluations with and without disturbances assess the broader benefit to unseen-task generalization (Section 6).

2 PRELIMINARIES

Markov Decision Process (MDP). Sequential decision-making tasks are modeled as MDPs (Kaelbling et al., 1996; Buşoniu et al., 2008). An episodic MDP $\tau \in \mathcal{M}$ is defined by the tuple $(\mathcal{S}, \mathcal{A}, P_\tau, R_\tau, \rho_\tau, H)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $P_\tau : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition function, $R_\tau : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\rho_\tau \in \Delta(\mathcal{S})$ is the initial state distribution, and H is the episode horizon. At each time step h, an agent selects an action $a_h \in \mathcal{A}$, receives a reward $r_h = R_\tau(s_h, a_h)$, and transitions to the next state s_{h+1} according to $P_\tau(s_h, a_h)$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps the current state to an action distribution. The agent's goal is to learn a policy $\pi^\star_\tau$ that maximizes the expected cumulative reward $J(\pi, \tau) = \mathbb{E}[\sum_{h=1}^{H} r_h \mid \pi, \tau]$. We use *nominal task* for the underlying MDP before an additional disturbance is applied; nominal tasks may differ in rewards, goals, or baseline dynamics. A nominal task may be seen or unseen in pretraining.

ICRL with Supervised Pretraining. We build on the supervised pretraining approach of Decision-Pretrained Transformer (DPT) (Lee et al., 2023). During pretraining, tasks $\{\tau^i\}_{i=1}^m$ are drawn from a task distribution p_τ over the MDP space $\mathcal{M}$. For each task τ^i, a behavioral policy collects a context dataset $\xi^i = \{(s^i_h, a^i_h, s^i_{h+1}, r^i_h)\}_h$, where a^i_h is selected by that policy. We also choose a query state $s^i_{\text{query}} \in \mathcal{S}$ and sample its optimal action $a^\star_i$ from the expert policy $\pi^\star_{\tau^i}(s^i_{\text{query}})$. The behavioral policy thus determines the experience available in context, while the expert provides the action target that the model learns to predict. The complete pretraining dataset is $\mathcal{D}_{pre} = \{\xi^i, s^i_{\text{query}}, a^\star_i\}_{i=1}^m$. Writing T_θ for the complete transformer-based policy, with θ collectively denoting its policy parameters, we train it to predict the expert action conditioned on the query state and context:

$$\min_\theta \frac{1}{m} \sum_{i=1}^{m} - \log T_\theta \left(a^\star_i | s^i_{\text{query}}, \xi^i\right). \tag{1}$$

ICRL Deployment. The pretrained policy T_θ can be deployed in both **offline** and **online** settings on an unseen task τ sampled from p_τ. In offline deployment, a behavioral policy first collects a context dataset ξ_{off} from τ. Once this dataset is available, the model selects actions using $T_\theta(\cdot|s_h, \xi_{\text{off}})$, adapting to the task through the supplied context. Standard online deployment instead begins with an empty dataset ξ_{on}. At each episode, the model follows $T_\theta(\cdot|s_h, \xi_{\text{on}})$ to collect a trajectory $\zeta = \{s_1, a_1, r_1, \ldots, s_H, a_H, r_H\}$, which is then appended to the context for subsequent episodes. The distinction is therefore whether the context is supplied beforehand or accumulated through interaction; in both cases, the pretrained model parameters remain fixed. Algorithm 3 in Appendix A gives these standard protocols. To use this mechanism for robust control, we must specify both the experience from which adaptation is learned and the controller that supplies its initial context.

3 FROM ADVERSARIAL TRAINING TO TAILORED CONTROL

We first consider the classical robust RL setting around a fixed nominal task. BRIDGE aims to reduce conservatism through a model capable of activating different tailored behaviors for encountered disturbances through context, without parameter updates. In comparison, canonical robust RL seeks a policy with strong worst-case return through max-min optimization (Moos et al., 2022):

$$\pi^{\text{rob}} \in \arg\max_\pi \min_{\tau \in \mathcal{U}} J(\pi, \tau), \tag{2}$$

where $J(\pi, \tau)$ denotes the expected return of policy $\pi(a|s)$ in environment τ. Here, $\mathcal{U}$ is the *uncertainty set* of perturbed MDPs around the fixed nominal task, and $\tau \in \mathcal{U}$ denotes one such

environment. The variations may arise from adversarial external forces (Pinto et al., 2017) or changes in physical constants such as friction and mass (Tanabe et al., 2022). Worst-case optimization in equation 2 often leads to conservative behavior when a fixed policy must hedge across these environments (Dong et al., 2025b). To this end, BRIDGE utilizes the optimization process itself to learn a deployment context-dependent response, using its intermediate experts to collect tailored behavior for *learning* and its final robust policies for the initial *deployment context*.

Central Design Question. To realize the BRIDGE principle, we pretrain an ICRL model T_θ to adapt across the perturbations within $\mathcal{U}$. The perturbed environments in $\mathcal{U}$ therefore play the role of the ICRL pretraining tasks introduced in Section 2, but specifying this space alone does not provide the experience needed for adaptation. The central question is thus *how to construct the pretraining task distribution p_τ and collect the corresponding expertise*.

Adversarial Training as a Curriculum. Our key insight is that *the max-min optimization process itself supplies a structured answer*. Conventional methods seek to solve equation 2 through iterative updates, traversing a sequence of policies and perturbations: $\pi^{(0)}, \tau^{(1)}, \pi^{(1)}, \ldots, \tau^{(T)}, \pi^{(T)}$, where each adversary identifies a weakness $\tau^{(t)}$, and the next policy $\pi^{(t)}$ learns to perform well under that disturbance. Along this trajectory, the intermediate policies $\pi^{(t)}$ become specialized experts for the corresponding disturbances $\tau^{(t)}$. Although robust RL deployment typically retains only the final robust policy $\pi^{(T)}$ and discards the intermediate adversaries and policies, ***CORE distills this valuable expertise generated throughout max-min training into T_θ***. Each training example pairs context from a perturbed environment $\tau^{(t)}$ with an expert action at a query state, supplied by the corresponding retained expert $\pi^{(t)}$. Learning these conditional responses makes tailored behavior available through in-context inference and provides experience from which to generalize to unseen disturbances.

Two-Phase Curriculum Design. Turning this distillation perspective into a pretraining pipeline requires both *coverage* of the uncertainty set and experience with *challenging perturbations*. Uniform sampling supplies initial variation across $\mathcal{U}$, but many sampled perturbations may induce similar dynamics and leave important policy weaknesses unexplored. CORE therefore combines uniform coverage with adversarial boundary expansion in the two phases of Algorithm 1 in Appendix A.

In **Phase I (Coverage)**, CORE uniformly samples perturbed environments $\tau \sim \text{Unif}(\mathcal{U})$. For each sampled environment, we train an expert policy $\pi^\star_\tau$ using standard RL and construct pretraining samples $(\xi_\tau, s_{\text{query}}, a^\star_\tau)$ following Section 2. The learning trajectories of $\pi^\star_\tau$ supply the context ξ_τ, while $a^\star_\tau \sim \pi^\star_\tau(\cdot \mid s_{\text{query}})$ supplies the action target for the query state. Repeating over N_I environments yields the initial dataset $\mathcal{D}_I$ and the expert policies to initiate the adversarial search.

In **Phase II (Boundary Expansion)**, CORE targets weaknesses of these experts to enrich the curriculum within the same uncertainty set $\mathcal{U}$. We initialize $\pi^{(0)}$ using an expert from Phase I, then search for the environment that minimizes its performance: $\tau^{(1)} \in \arg\min_{\tau \in \mathcal{U}} J(\pi^{(0)}, \tau)$. Having found this weakness, we train a new expert policy $\pi^{(1)} \in \arg\max_\pi J(\pi, \tau^{(1)})$ and use the resulting expert to generate additional pretraining data in $\tau^{(1)}$. Subsequent iterations repeat this logic, finding an adversary $\tau^{(t)}$ against the previous expert $\pi^{(t-1)}$ and training the next expert policy $\pi^{(t)}$. We perform N_{II} independent runs, each initialized by randomly selecting a Phase I expert and comprising T max-min steps. These runs can explore different challenging regions of $\mathcal{U}$.

Computation. Generating the curriculum requires repeated RL optimization before deployment. Phase I samples and independent Phase II runs can be processed *in parallel*, while updates within each run remain sequential. Deployment uses *fixed* parameters; Appendix C.3 discusses more details.

Robust Context Collection and Deployment. Distilling the curriculum gives T_θ access to tailored expertise, but using that expertise requires a context dataset ξ from the deployment environment. This creates a **cold-start dilemma**: *the agent must interact with a potentially disturbed environment before it has enough evidence to infer the disturbance*. We address it by retaining the final controllers $\pi^{(T)}$ from max-min training as the robust behavioral policies for initial context collection. Executing the same robust baseline retains its behavior during collection. After initial context collection, T_θ takes over and conditions its actions on the collected experience, with its parameters fixed. Section 5 quantifies when this tailored response improves worst-case return. Section 6 reports the deployment protocol and empirical gains over robust RL baselines both with and without adversarial forces; Appendix F separately ablates the curriculum and robust initialization (Table 6).

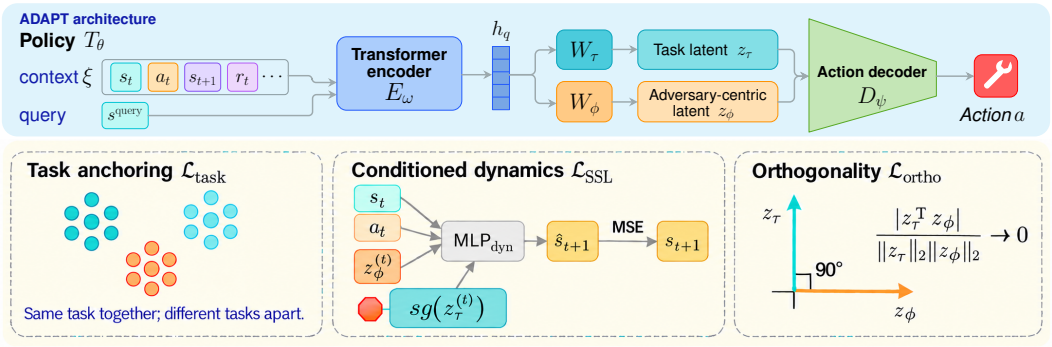


Figure 2: **ADAPT architecture.** Policy T_θ (top), including encoder E_ω , projection heads W_τ and W_ϕ, action decoder D_ψ, with auxiliary pretraining objectives (bottom) to facilitate task-disturbance information disentanglement. All policy parameters remain fixed at deployment.

4 Robust Generalization to Unseen Tasks Under Disturbance

Section 3 shows how ICRL benefits robust RL. We now take BRIDGE one step further, pursuing the reciprocal benefit: using adversarial curricula to strengthen ICRL's own generalization to unseen nominal tasks under deployment disturbances.

Pretraining and Task Inference under Disturbances. For each nominal training task τ from a shared family $\mathcal{M}$, we construct a CORE curriculum over its uncertainty set $\mathcal{U}(\tau)$. We pretrain a single policy on the *combined* curricula using expert action supervision; deployment then requires generalization to held-out nominal tasks from the same family without further max-min training. This creates a **joint inference challenge**. Around a fixed nominal task, context must reveal the disturbance; on an unseen task, it must also reveal what the agent needs to accomplish. *Observed changes may reflect a different task, a disturbance, or both, motivating separate task and adversary-centric representations*. Adversarial variations also enrich the pretraining distribution, with benefits extending to unseen tasks even without deployment disturbances, as shown in Section 6.

ADAPT Architecture. We address this challenge with the **A**dversary-aware **Du**A**l**-stream **P**olicy **T**ransformer (ADAPT). The complete policy T_θ uses a shared transformer encoder E_ω, where θ collects all policy parameters and ω denotes encoder parameters. The encoder processes transition context $\xi = \{(s_t, a_t, s_{t+1}, r_t)\}_t$ and the separate query state s_{query}, producing $h_q = E_\omega(\xi, s_{\text{query}})$. Projection heads produce the task latent $z_\tau = W_\tau h_q$ and adversary-centric latent $z_\phi = W_\phi h_q$; the action decoder combines these late-branching representations as $a = D_\psi(z_\tau, z_\phi)$. We train the complete policy to predict expert actions using the supervised objective of Section 2:

$$\mathcal{L}_{SL} = -\mathbb{E}_{\mathcal{D}_{multi}}[\log T_\theta(a^\star \mid \xi, s_{\text{query}})]\,, \tag{3}$$

where the expectation is over $(\xi, s_{\text{query}}, a^\star) \sim \mathcal{D}_{multi}$, the combined curricula of the nominal training tasks. Separate heads alone do not determine the latents' roles; we guide them with three auxiliary pretraining objectives: *(i) task anchoring, (ii) conditioned dynamics prediction, and (iii) geometric orthogonality*, illustrated in Figure 2. We discuss each of them in detail below, and Algorithm 2 in Appendix A gives the full pretraining procedure.

Task Anchoring across Disturbances. We encourage the task latent to retain information shared across perturbed realizations of the same nominal task. Using available training-task labels, the InfoNCE loss (Oord et al., 2018) pairs an anchor context i with a positive context i^+ from the same nominal task under a different perturbation and negatives from different nominal tasks:

$$\mathcal{L}_{task} = -\mathbb{E}_i\left[\log \frac{\exp\left(z_{\tau,i}^\top z_{\tau,i^+}/\kappa\right)}{\sum_{j\in\mathcal{J}_i}\exp\left(z_{\tau,i}^\top z_{\tau,j}/\kappa\right)}\right], \tag{4}$$

where $\mathcal{J}_i$ contains the positive and negative candidates and $\kappa > 0$ is the temperature. This encourages stability across disturbances while preserving task distinctions needed for decision-making. These labels supervise representation learning on the training tasks; at deployment, task information must be inferred from context for unseen nominal tasks.

Adversary Representation through Conditioned Dynamics. Nominal task labels support contrastive pairing, but CORE does not generally reuse adversaries across tasks. We therefore guide the adversary latent with self-supervised dynamics prediction conditioned on the task latent, following Algorithm 2. Writing $z_\tau^{(t)}$ and $z_\phi^{(t)}$ for the transition representations used by the dynamics head, we predict the next state from the current state, action, and latents:

$$\mathcal{L}_{SSL} = \mathbb{E}\left[\|s_{t+1} - \widehat{s}_{t+1}\|_2^2\right]; \quad \widehat{s}_{t+1} = \mathrm{MLP}_{dyn}\Big(s_t, a_t, \mathrm{sg}(z_\tau^{(t)}), z_\phi^{(t)}\Big). \tag{5}$$

The *stop-gradient* operator $\mathrm{sg}(\cdot)$ detaches $z_\tau^{(t)}$ from the dynamics loss, so this loss gives no gradient to the task projection head. Gradients through $z_\phi^{(t)}$ still update the shared encoder and can therefore change the task latent indirectly. Conditioned on the task latent, the objective encourages the adversary latent to encode additional information useful for next-state prediction.

Geometric Orthogonality Constraint. The shared early attention layers can produce overlapping latent features despite these objectives. We therefore discourage alignment between the two latent vectors with a geometric orthogonality penalty:

$$\mathcal{L}_{ortho} = \mathbb{E}\left[\frac{|z_\tau^\top z_\phi|}{\|z_\tau\|_2 \|z_\phi\|_2}\right]. \tag{6}$$

Joint Pretraining and Frozen Deployment. The action objective trains the representations to support decision-making jointly, while the auxiliary objectives encourage complementary roles. We optimize all model parameters offline using

$$\mathcal{L}_{total} = \mathcal{L}_{SL} + \lambda_1 \mathcal{L}_{task} + \lambda_2 \mathcal{L}_{SSL} + \lambda_3 \mathcal{L}_{ortho}, \qquad \lambda_1, \lambda_2, \lambda_3 \geq 0. \tag{7}$$

Section 6 evaluates ADAPT on unseen tasks across benchmarks with and without disturbances, demonstrating its competitive performance. To assess the contributions of its components, Section 6.2 reports representation ablations and latent probes.

5 Performance Guarantees for Dynamic Generalization

In this section, we answer the central performance question through analysis: *when does in-context generalization realize the advantage of tailored control, and what does the gain look like*? In particular, we show that BRIDGE can improve worst-case return when the advantage of tailored control exceeds the error in recovering it from expert supervision, deployment context, and finite pretraining. We quantify these three sources of error and extend the guarantee to fresh nominal tasks.

Consider horizon-H MDPs with rewards in $[0, 1]$. Let $J^*(\psi)$ be the disturbance-informed optimal return and $\bar{J}_\theta(\psi)$ the post-context return of the frozen policy, averaged over the actual collector's context. Parameters and context remain fixed within each evaluation episode. For a conventional policy class Π_0 without this additional context, define

$$V^{\mathrm{rob}} = \sup_{\pi \in \Pi_0} \inf_{\psi \in \Psi} J(\pi, \psi), \qquad \Delta_{expert} = \inf_{\psi \in \Psi} J^*(\psi) - V^{\mathrm{rob}} \geq 0. \tag{8}$$

Theorem 5.1 (Performance gain of BRIDGE)**.** *Under the sampling and prediction conditions of Appendix G.1, suppose joint training-to-deployment coverage is bounded by C, the realizable predictor class has size K, and pretraining uses M independent collections. On the compact disturbance space, assume each radius-r neighborhood has curriculum probability at least cr^d for $0 < r \leq r_0$, and context-averaged regret is L-Lipschitz. Then, with probability at least $1 - \delta$,*

$$\underbrace{\inf_{\psi} \bar{J}_{\widehat{\theta}}(\psi) - V^{\mathrm{rob}}}_{\textit{worst-case performance gain}} \geq \underbrace{\Delta_{expert}}_{\textit{tailored-control advantage}} - \mathcal{B}_d(\varepsilon_{N,M}), \tag{9}$$

where the average regret is bounded by

$$\varepsilon_{N,M} = \underbrace{\epsilon_E}_{\textit{expert error}} + H^2 \left[\underbrace{\sqrt{\frac{C\mathcal{I}_N}{2}}}_{\textit{context uncertainty}} + \underbrace{\sqrt{C\zeta_M + \frac{2C\log(K/\delta)}{M}}}_{\textit{pretraining error}} \right]. \tag{10}$$

Here ϵ_E is average expert suboptimality, $\mathcal{I}_N = I(A^\star; \psi \mid \xi, s, h)$ is residual expert-action uncertainty with context size N, and ζ_M is empirical action-loss optimization error. The transform $\mathcal{B}_d(u) = \min\{H, \inf_{0<r\le r_0}[u/(cr^d) + Lr]\}$ converts average error into worst-case error.

When contextual specialization pays off. Equation 9 guarantees strict improvement when $\mathcal{B}_d(\varepsilon_{N,M}) < \Delta_{expert}$. Under the informative-context conditions of Appendix G.6, $\mathcal{I}_N = O(N^{-1})$, giving $\varepsilon_{N,M} \le \epsilon_E + H^2\sqrt{C\zeta_M} + \widetilde{O}(N^{-1/2} + M^{-1/2})$ for fixed problem and class constants. Here d describes local coverage of the disturbance family, not necessarily an adversary network's parameter count. For families indexed by a few physical variables, such as mass and friction, d can be small under the stated coverage and regret-regularity conditions. Appendix G.5 explains this interpretation and proves that the conversion exponent $1/(d+1)$ cannot be improved without stronger assumptions.

Scale of Gain and Empirical Validation. The available advantage Δ_{expert} can be substantial when disturbance-specific experts maintain high returns but a conventional policy must compromise between conflicting control strategies. The experiments in Section 6 examine whether BRIDGE realizes this advantage in practice. Indeed, we observe substantial performance gains in robust control and unseen-task generalization, both with and without deployment disturbances.

Generalization to Unseen Tasks. Corollary G.10 in Appendix G.9 extends the guarantee to the worst disturbance per task, averaged over fresh nominal tasks from the covered population. Context must now resolve action-relevant uncertainty about both the task and its disturbance. The information decomposition in equation 52 formalizes this joint inference challenge, connecting the analysis to ADAPT's complementary representations. The extension retains the disturbance-space exponent under uniform within-task coverage and regularity. Further appendix results cover the full collection-and-deployment horizon (Proposition G.13), approximate predictor classes (Lemma G.5), continuous actions (Appendix G.7), and representation and decoding error (Appendix G.10).

6 Experiments

We evaluate BRIDGE's reciprocal strengths: in-context adaptation for robust control (MuJoCo), and generalization to unseen tasks *with and without disturbances* (Dark Room and Meta-World). Ablations then examine the curriculum, initial context, and representations.

Evaluation protocol. Following the offline/online settings in Section 2, we report returns over 10 seeds; shaded regions and table intervals denote the standard error of the mean (SEM). Online deployment begins with $K_0 = 5$ episodes collected by uniformly sampling from the robust endpoint ensemble $\Pi^{\mathrm{rob}} = \{\pi_i^{(T)}\}_{i=1}^{N_{\mathrm{II}}}$. Frozen ADAPT then takes over and appends subsequent trajectories to $\mathcal{D}_{\mathrm{on}}$. Appendices C and D give model settings, computation, datasets, and full evaluation details.

Baselines. We compare with **ICRL baselines**: Prompt-DT (Xu et al., 2022), ARDT (Tang et al., 2024), Mix-DPT (Wang et al., 2024), DPT (Lee et al., 2023), and an adapted BATI (Ma et al., 2025); **robust RL baselines**: Domain Randomization (DR) (Tobin et al., 2017), Max-Min TD3 (M2TD3) (Tanabe et al., 2022), RARL (Pinto et al., 2017), and QARL (Reddi et al., 2024); and **adaptive robust RL baselines**: PROTECTED (Liu et al., 2024) and SeCBAD (Chen et al., 2022). Applicable baselines vary by domain; their implementation details are in Appendix C.

6.1 Robust Control and Generalization to Unseen Tasks

MuJoCo: Adversarial Forces and Unseen Dynamics. We evaluate 11 continuous-control tasks from the DeepMind Control Suite (Tunyasuvunakool et al., 2020), first at nominal physical parameters with and without adversarial forces, and then under unseen dynamics (Appendix E).

At the nominal physical parameters used to train the standard RL baselines, ADAPT achieves higher median returns than RARL and QARL on Quadruped-Run and Walker-Run both **with and without adversarial forces** (Figure 8). This supports improved nominal performance and robustness under the evaluated disturbances. ADAPT also improves robustness over DPT, particularly on Quadruped-Run.

To evaluate **unseen physical parameters**, we vary two parameters over an 8×8 grid, while ICRL pretraining covers only the central 4×4 region. ADAPT generalizes beyond this region, with clear gains in Walker-Run and Cheetah-Run (Figures 12b and 13a). Appendix E gives complete results and protocols. We next examine unseen goals and manipulation configurations.

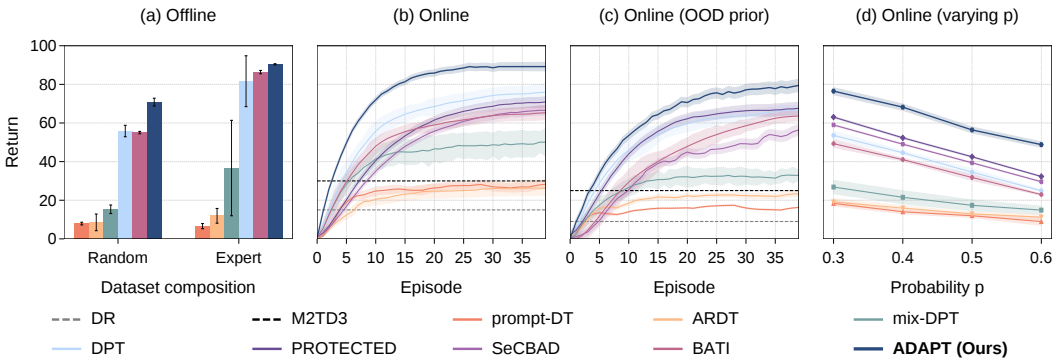


Figure 3: Held-out Dark Room goals: (a) offline; (b) undisturbed online; (c) online with OOD disturbance priors ($p = 0.2$); (d) varying disturbance probability p after 40 episodes.

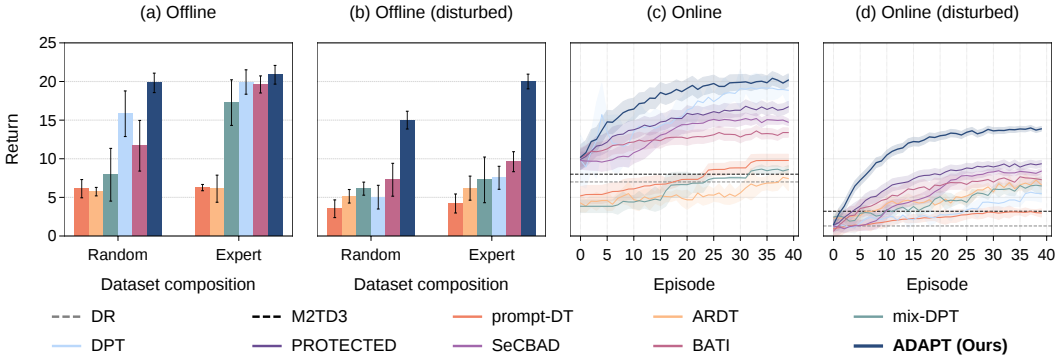


Figure 4: Generalization to held-out Meta-World tasks with and without observation disturbances.

Dark Room: Unseen Goals and Action Disturbances. Dark Room is a sparse-reward navigation task with hidden goals on a 10×10 grid. Following Lee et al. (2023), we pretrain on 80 goals and evaluate on 20 **held-out test goals**. A Dirichlet-based disturbance can override the intended action, requiring context to reveal both the unseen goal and disturbed action execution (Appendix D.1).

ADAPT performs well on held-out goals with either random or expert context (Figure 3a), showing that adaptation can use experience of different behavioral quality. In online evaluation **without disturbances**, it also improves sample efficiency and final return over 40 episodes (Figure 3b).

With an **unseen Dirichlet disturbance prior** at takeover probability $p = 0.2$, ADAPT's advantage becomes more pronounced (Figure 3c). Varying the takeover probability further tests robustness: after 40 episodes, ADAPT maintains strong returns even at $p = 0.6$ (Figure 3d). Thus, the unseen-goal gains persist under disturbances to action execution.

Meta-World: Generalization to Unseen Tasks under Observation Disturbances. To test whether these gains extend to continuous robotic manipulation, we use Meta-World ML1 pick-place, pretraining on 15 configurations and evaluating on 5 **held-out configurations** with novel object and target positions. Disturbed evaluations use a fixed-weight neural-network adversary applying bounded **observation perturbations** ($|\phi^{i,k}| \leq 1$); Appendix D gives the details.

Figure 4 evaluates supplied random/expert context and online adaptation, both **with and without disturbances**. Here, the agent must infer the new configuration from context that may itself be disturbed. Although observation disturbances reduce returns, ADAPT maintains stronger performance, including against PROTECTED and SeCBAD in online adaptation. Together, these evaluations support retaining useful task information while adapting control to disturbed execution.

6.2 Ablations and Representation Analysis

Adversarial curriculum and robust context collection. With the ADAPT architecture fixed, spending the full Phase I + II budget on Phase I yields an Episode 40 return of 7.4, compared with 13.1 for the two-phase curriculum (Table 1). Thus, the gains from adversarial search persist under the same budget. With robust initialization, the corresponding returns are 9.3 and 13.9; initialization also raises the full curriculum's Episode 5 return from 3.8 to 6.5. Appendix F reports additional episode-level results.

Table 1: CORE on held-out disturbed Meta-World tasks (10-seed mean return ± SEM).

Setup	Ep. 5	Ep. 40
Phase I	1.6 ± 0.4	6.5 ± 0.5
Phase I (full budget)	1.9 ± 0.5	7.4 ± 0.6
+ robust init.	3.1 ± 0.3	9.3 ± 0.4
Phase I + II	3.8 ± 0.6	13.1 ± 0.5
+ robust init.	$\mathbf{6.5 \pm 0.4}$	$\mathbf{13.9 \pm 0.2}$

Task and adversary representations. Table 2 shows gains across Meta-World and three MuJoCo tasks as we add dual heads with task anchoring, conditioned dynamics prediction, and orthogonality. Relative to the single-stream baseline, full ADAPT increases Meta-World return from 6.1 to 13.9 and Quadruped-Run return from 157.2 to 186.0.

Table 2: Cumulative ADAPT component ablation on disturbed Meta-World (Ep. 40) and MuJoCo.

Variant	Meta-World	Quadruped-Run	Walker-Run	Reacher-Hard
Single-stream baseline	6.1 ± 0.5	157.2 ± 8.2	117.3 ± 3.5	169.2 ± 6.9
Dual-stream + $\mathcal{L}_{task}$	8.7 ± 0.6	166.2 ± 4.8	125.4 ± 4.6	173.8 ± 4.1
+ $\mathcal{L}_{SSL}$	10.6 ± 0.4	181.2 ± 3.4	130.7 ± 4.1	180.8 ± 3.2
+ $\mathcal{L}_{ortho}$ (full ADAPT)	$\mathbf{13.9 \pm 0.2}$	$\mathbf{186.0 \pm 2.6}$	$\mathbf{140.4 \pm 6.8}$	$\mathbf{186.6 \pm 2.8}$

Linear Probing of Learned Representations. Linear probes assess task and adversary information in frozen latents (Table 7; setup in Appendix F). For task identity, the task latent achieves 91% accuracy, compared with 35% for the adversary latent and 58% for the single-stream feature. For adversary identity, the adversary latent achieves 85%, compared with 26% for the task latent and 51% for the single-stream feature. Within-target comparisons support complementary latent roles.

7 Related Work

Robust RL studies worst-case control in uncertain MDPs (Nilim & Ghaoui, 2005; Iyengar, 2005), with methods addressing uncertain dynamics (Tanabe et al., 2022; Rajeswaran et al., 2017) and adversarial perturbations (Pinto et al., 2017; Tessler et al., 2019; Zhang et al., 2020; Reddi et al., 2024). ICRL learns adaptation from learning histories in Algorithm Distillation or expert-action prediction in DPT (Laskin et al., 2023; Lee et al., 2023). History-conditioned controllers learn latent adaptation or recurrent robust policies (Kumar et al., 2021; Zhang et al., 2021; Hsu et al., 2024), while ARDT uses minimax-return conditioning (Tang et al., 2024). BRIDGE instead distills intermediate adversarial experts and reuses the final controller for context, connecting robust control to unseen-task generalization. Appendix B provides the full discussion.

8 Discussion and Limitations

BRIDGE develops robustness as dynamic generalization without parameter updates. ICRL makes tailored control available from deployment context, while adversarial robust RL strengthens ICRL's generalization to unseen tasks. CORE incurs additional computation for pretraining data collection. When nominal training tasks are limited, adversarial variations can make this a useful *tradeoff* for improving generalization and robustness; this offline cost can be *amortized* across deployments of the same pretrained model. The benefits depend on informative context and incur inference costs; guarantees additionally require deployment coverage and an environment law fixed over collection and evaluation. Robust initial collection preserves the chosen controller's behavior, but obtaining useful context can delay adaptation. Extending the framework to rapidly changing environments and reducing its context and computation requirements remain important directions for further work.

AI USE STATEMENT

Generative AI tools were employed exclusively for manuscript editing, stylistic refinement, layout structuring of figures, and generating the robot and icon cartoons in Figure 1. All theoretical formulations, proofs, algorithmic designs, software implementations, and experimental evaluations were conceived, verified, and executed directly by the authors.

REPRODUCIBILITY STATEMENT

The benchmark environments and pretraining-data construction are described in Section 6 and Appendix D. Appendix C provides baseline implementations, model architecture, and hyperparameter settings; Appendix A gives the algorithms. Complete proofs and the assumptions underlying the analysis are included in Appendix G.

REFERENCES

Andrew Gehret Barto, Richard S Sutton, and CJCH Watkins. *Learning and sequential decision making*, volume 89. University of Massachusetts Amherst, MA, 1989.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Xiaoyu Chen, Xiangming Zhu, Yufeng Zheng, Pushi Zhang, Li Zhao, Wenxue Cheng, Peng Cheng, Yongqiang Xiong, Tao Qin, Jianyu Chen, and Tie-Yan Liu. An adaptive deep RL method for non-stationary environments with piecewise stable context. In *Advances in Neural Information Processing Systems*, 2022.

Juncheng Dong, Moyang Guo, Ethan X Fang, Zhuoran Yang, and Vahid Tarokh. In-context reinforcement learning from suboptimal historical data. In *International conference on machine learning*. PMLR, 2025a.

Juncheng Dong, Hao-Lun Hsu, Qitong Gao, Vahid Tarokh, and Miroslav Pajic. Variational adversarial training towards policies with improved robustness. In *Artificial intelligence and statistics*. PMLR, 2025b.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Hao-Lun Hsu and Miroslav Pajic. Robust exploration with adversary via Langevin Monte Carlo. In Alessandro Abate, Mark Cannon, Kostas Margellos, and Antonis Papachristodoulou (eds.), *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1592–1605. PMLR, 15–17 Jul 2024.

Hao-Lun Hsu, Haocheng Meng, Shaocheng Luo, Juncheng Dong, Vahid Tarokh, and Miroslav Pajic. Reforma: Robust reinforcement learning via adaptive adversary for drones flying under disturbances. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5169–5175, 2024.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems*, 2021. URL https://arxiv.org/abs/2107.04034.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *International Conference on Learning Representations*, 2023.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.

Xiangyu Liu, Chenghao Deng, Yanchao Sun, Yongyuan Liang, and Furong Huang. Beyond worst-case attacks: Robust RL with adaptive defense via non-dominated policies. In *The Twelfth International Conference on Learning Representations*, 2024.

Long Ma, Fangwei Zhong, and Yizhou Wang. Behavior-agnostic task inference for robust offline in-context reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025.

Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International conference on machine learning*, pp. 2817–2826. PMLR, 2017.

Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. EPOpt: Learning robust neural network policies using model ensembles. *International Conference on Learning Representations*, 2017.

Aryaman Reddi, Maximilian Tölle, Jan Peters, Georgia Chalvatzaki, and Carlo D'Eramo. Robust adversarial reinforcement learning via bounded rationality curricula. *International Conference on Learning Representations*, 2024.

Paulius Sasnauskas, Yiğit Yalın, and Goran Radanović. Robust in-context reinforcement learning under reward poisoning attacks. In *International Conference on Machine Learning*, 2026. URL `https://arxiv.org/abs/2506.06891`.

Takumi Tanabe, Rei Sato, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. Max-min off-policy actor-critic method focusing on worst-case robustness to model misspecification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Xiaohang Tang, Afonso Marques, Parameswaran Kamalaruban, and Ilija Bogunovic. Adversarially robust decision transformer. *Advances in neural information processing systems*, 2024.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224. PMLR, 2019.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: software and tasks for continuous control. *Software Impacts*, 6, 2020.

Hanzhao Wang, Yu Pan, Fupeng Sun, Shang Liu, Kalyan Talluri, Guanting Chen, and Xiaocheng Li. Understanding the training and generalization of pretrained transformer for sequential decision making. *arXiv preprint arXiv:2405.14219*, 2024.

Puyue Wang, Jiawei Hu, Yan Gao, Junyan Wang, Yu Zhang, Gillian Dobbie, Tao Gu, Wafa Johal, Ting Dang, and Hong Jia. HoRD: Robust humanoid control via history-conditioned reinforcement learning and online distillation. *arXiv preprint arXiv:2602.04412*, 2026. URL https://arxiv.org/abs/2602.04412v1.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pp. 24631–24645. PMLR, 2022.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*. PMLR, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.08452.

Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, and Sergey Kolesnikov. Emergence of in-context reinforcement learning from noise distillation. In *International conference on machine learning*. PMLR, 2024.

A PSEUDOCODE

Algorithm 1 constructs the adversarial curriculum for in-context robust control. Applying that construction across nominal training tasks yields $\mathcal{D}_{multi}$ for ADAPT, whose pretraining procedure is given in Algorithm 2. The ADAPT objectives are defined in Section 4. Algorithm 3 gives the standard offline and online ICRL protocols of Section 2; robust context initialization is discussed separately in Section 3.

Algorithm 1 Curriculum of Optimal Robust Experts (CORE)

Require: Uncertainty set $\mathcal{U}$, Phase I samples N_I, Phase II runs N_{II}, max-min steps T
Ensure: Pretraining dataset $\mathcal{D}$, trained sequence model T_θ, robust ensemble Π^{rob}

Initialize dataset $\mathcal{D} \leftarrow \emptyset$ and robust ensemble $\Pi^{\mathrm{rob}} \leftarrow \emptyset$
% Phase I: Uniform Coverage
for $i = 1$ **to** N_I **in parallel do**
    Sample perturbed environment $\tau \sim \mathrm{Unif}(\mathcal{U})$
    Train expert policy $\pi^\star_\tau \in \arg\max_\pi J(\pi, \tau)$
    Collect trajectories ξ_τ using $\pi^\star_\tau$ in environment τ
    Choose query states s_{query} and obtain expert targets $a^\star_\tau \sim \pi^\star_\tau(\cdot \mid s_{\mathrm{query}})$
    Add samples $(\xi_\tau, s_{\mathrm{query}}, a^\star_\tau)$ to $\mathcal{D}$, following Section 2
end for

% Phase II: Boundary Expansion via Max-Min Optimization
for $i = 1$ **to** N_{II} **in parallel do**
    Initialize $\pi^{(0)}$ by uniformly sampling from Phase I experts $\{\pi^\star_\tau\}$
    for $t = 1$ **to** T **do**
        % Find the most challenging adversary for the current expert
        $\tau^{(t)} \leftarrow \arg\min_{\tau \in \mathcal{U}} J(\pi^{(t-1)}, \tau)$
        % Train a new expert policy
        $\pi^{(t)} \leftarrow \arg\max_\pi J(\pi, \tau^{(t)})$
        % Collect context trajectories
        Collect trajectories $\xi^{(t)}$ using $\pi^{(t)}$ in environment $\tau^{(t)}$
        Choose query states s_{query} and obtain expert targets $a^{\star,(t)} \sim \pi^{(t)}(\cdot \mid s_{\mathrm{query}})$
        Add samples $(\xi^{(t)}, s_{\mathrm{query}}, a^{\star,(t)})$ to $\mathcal{D}$, following Section 2
    end for
    $\Pi^{\mathrm{rob}} \leftarrow \Pi^{\mathrm{rob}} \cup \{\pi^{(T)}\}$ *% Retain endpoint for context collection*
end for

% Pretraining
Train ICRL sequence model T_θ on dataset $\mathcal{D}$ via supervised pretraining in equation 1.
return $\mathcal{D}$, T_θ, Π^{rob}

Algorithm 2 ADAPT: Adversary-aware DuAl-stream Policy Transformer

Require: Pretraining dataset $\mathcal{D}_{multi}$, transformer encoder E_ω, projection heads W_τ, W_ϕ, decoder D_ψ, head MLP_{dyn}
Require: Hyperparameters: Loss weights $\lambda_1, \lambda_2, \lambda_3$, InfoNCE temperature κ

▷ — *Forward Pass Procedure* —

procedure FORWARDPASS(ξ, s_{query})
  $h_q \leftarrow E_\omega(\xi, s_{\text{query}})$ ▷ Sequence representation at query state
  $z_\tau \leftarrow W_\tau h_q$ ▷ Task latent
  $z_\phi \leftarrow W_\phi h_q$ ▷ Adversary-centric latent
  $a \leftarrow D_\psi(z_\tau, z_\phi)$ ▷ Robust action prediction
  return a, z_τ, z_ϕ, h_q
end procedure

▷ — *Pretraining Procedure* —

while not converged **do**
  Sample batch $(\xi, s_{\text{query}}, a^\star, \tau)$ from $\mathcal{D}_{multi}$
  $a, z_\tau, z_\phi, h_q \leftarrow$ FORWARDPASS(ξ, s_{query})
  ▷ **Joint Action and Representation Objectives**
  Compute the supervised action loss $\mathcal{L}_{SL}$ using equation 3
  Form nominal-task positive and negative pairs; compute $\mathcal{L}_{task}$ using equation 4
  $\hat{s}_{t+1} \leftarrow \text{MLP}_{dyn}(s_t, a_t, \text{sg}(z_\tau^{(t)}), z_\phi^{(t)})$ ▷ Condition on task stop-gradient
  $\mathcal{L}_{SSL} = \mathbb{E}[\|s_{t+1} - \hat{s}_{t+1}\|_2^2]$
  $\mathcal{L}_{ortho} = \mathbb{E}[|\frac{z_\tau^\top z_\phi}{\|z_\tau\|\|z_\phi\|}|]$
  $\mathcal{L}_{total} = \mathcal{L}_{SL} + \lambda_1 \mathcal{L}_{task} + \lambda_2 \mathcal{L}_{SSL} + \lambda_3 \mathcal{L}_{ortho}$
  Update all parameters using $\nabla \mathcal{L}_{total}$
end while

Algorithm 3 Standard Offline and Online ICRL Deployment

Input: Pretrained complete policy T_θ; Horizon of episodes H; Number of episodes N for online testing; Offline dataset $\xi_{\text{off}} = \{(s_h, a_h, s_{h+1}, r_h)\}_h$, consisting of transitions collected by a behavioral policy.
// Offline Testing
for every time step $h \in \{1, \ldots, H\}$ **do**
  Observe state s_h
  Sample action with T_θ:

$$a_h \sim T_\theta\left(\cdot | s_h, \xi_{\text{off}}\right)$$

  Collect reward r_h
end for
// Online Testing
Initialize an empty online data buffer $\xi_{\text{on}} = \{\}$
for every online trial $n \in \{1, \ldots, N\}$ **do**
  for every time step $h \in \{1, \ldots, H\}$ **do**
    Observe state s_h
    Sample action with T_θ:

$$a_h \sim T_\theta\left(\cdot | s_h, \xi_{\text{on}}\right)$$

    Collect reward r_h
  end for
  Append the collected transitions $\{(s_h, a_h, s_{h+1}, r_h)\}_h$ into ξ_{on}
end for

B Related Work

Adaptive robustness with fixed parameters. Robust MDP formulations optimize performance over an uncertainty set (Nilim & Ghaoui, 2005; Iyengar, 2005). Deep implementations use parametric uncertainty (Tanabe et al., 2022; Rajeswaran et al., 2017) or adversarial perturbations (Pinto et al., 2017; Tessler et al., 2019; Zhang et al., 2020; Reddi et al., 2024). These objectives do not require a history-independent policy. RMA (Kumar et al., 2021) trains a privileged environment encoder and controller, then learns to infer the environment latent from recent state–action history. ATLA (Zhang et al., 2021) alternates protagonist and observation-adversary training and demonstrates the benefits of recurrent policies. REFORMA (Hsu et al., 2024) combines adversarial training with latent adaptation using histories of states, intended actions, and attacked actions. BRIDGE builds on fixed-parameter adaptation through a complementary construction: intermediate adversarial experts supply pretraining supervision, and a retained robust controller collects initial context. The collection guarantee is relative to that controller; subsequent improvement depends on context information and prediction quality (Section 5).

ICRL and the object of distillation. Offline RL learns decision rules from fixed datasets (Levine et al., 2020; Prudencio et al., 2023); transformer approaches include return-conditioned Decision Transformer (Chen et al., 2021) and demonstration-conditioned Prompt-DT (Xu et al., 2022). Algorithm Distillation (AD) (Laskin et al., 2023) models multi-episode learning histories, allowing policy improvement through context with fixed network parameters. DPT (Lee et al., 2023) instead predicts an expert's action at a query state from a separately collected context dataset. Its evaluations already include held-out goals and unseen action permutations. CORE builds on these foundations by supplying the specialized expert policies generated during max-min optimization: each sample pairs context from a perturbed environment with the associated checkpoint's expert action. This supervision does not require reproducing the ordered optimization trace used by AD. Other ICRL methods relax the quality or availability of expert trajectories (Dong et al., 2025a; Zisman et al., 2024).

History-conditioned transformer distillation. HoRD (Wang et al., 2026) trains a history-conditioned teacher under domain randomization and distills its actions into a transformer using simulation rollouts. Its history module infers execution dynamics, while motion commands specify the desired task. Its online distillation concerns student training rather than deployment-time weight updates. CORE instead supplies a sequence of adversarially generated experts, and ADAPT addresses deployment in which context must also identify an unseen nominal task.

Different robustness targets in sequence models. ARDT (Tang et al., 2024) estimates minimax returns through expectile regression, relabels offline trajectories, and trains a return-conditioned transformer. Its supervision therefore targets worst-case returns rather than CORE's specialist actions. BATI (Ma et al., 2025) addresses shifts in context-collection behavior using a likelihood model of rewards and next states, selecting task embeddings through a finite candidate search. AT-DPT (Sasnauskas et al., 2026) addresses reward poisoning: attackers learn to corrupt the DPT's context while the DPT learns to predict clean-task oracle actions. Its adversaries train directly against the transformer, whereas CORE generates expert experience through adversarial RL before distillation. Our task-family setting evaluates generalization to held-out nominal tasks under environmental disturbances (Section 4), connected to robust control around a fixed task by the shared curriculum construction.

C Implementation Details

Evaluation seeds. Experiments report returns over 10 random seeds. Seeds control parameter initialization and pretraining-dataset sampling for offline methods, and network and environment initialization for online RL baselines. Shaded regions and table intervals denote the standard error of the mean across seeds.

C.1 Algorithms

Soft Actor-Critic (SAC) (Haarnoja et al., 2018). SAC is an off-policy RL algorithm that balances expected return and action entropy. An actor selects actions, and two critics estimate Q-values.

Its objective is $J(\pi) = \sum_h \mathbb{E}_{(s_h,a_h)\sim D}[Q(s_h, a_h) - \alpha \log \pi(a_h|s_h)]$, where α controls entropy regularization. We use the implementation without adversarial training from Reddi et al. (2024).

Decision-Pretrained Transformer (DPT) (Lee et al., 2023). DPT uses supervised pretraining across tasks to learn context-conditioned decision-making. For each task τ_i, a behavioral policy collects a context dataset D^i of prior transitions, and an expert supplies the optimal action for a query state s^i_{query}. The transformer is trained to predict this action conditioned on both the query state and the context, minimizing the expected prediction loss over the pretraining task distribution. Separating behavioral context from expert action labels allows the model to learn useful responses even when the context trajectories are suboptimal. This is the supervised ICRL construction used by BRIDGE in Section 2; CORE determines the adversarial tasks and expertise included in its pretraining experience.

Mixed Decision-Pretrained Transformer (mix-DPT) (Wang et al., 2024). The mix-DPT framework extends DPT by splitting the learning process into two phases: the early training phase and the mixed training phase, addressing the out-of-distribution (OOD) issue between training and testing. During the early training phase, data are generated using a pre-specified decision function f, such as a random policy for the Dark Room task or SAC-trained policies for Meta-World and MuJoCo control. In the mixed training phase, data are generated using both the function f and the current DPT model, with the proportion controlled by a hyper-parameter κ. In the experiments conducted by Wang et al. (2024), historical trajectories include the optimal actions for each time step. To ensure a fair comparison, we follow their setup but use only a query state and the corresponding optimal action, as done in DPT and ADAPT. For our experiments, we follow the parameter choice $\kappa = 1/3$ from Wang et al. (2024), with the total number of training iterations set to 40% of the overall training iterations.

Behavior-agnostic Task Inference (BATI) (Ma et al., 2025). Published BATI infers tasks using the conditional likelihood of both rewards and next states given a task embedding. It selects among sampled embedding candidates at deployment, with learned networks fixed, and uses an IQL policy. Our experiments use a modified implementation: observations include a sliding history window, the dynamics model predicts the next augmented state, and the task latent z is optimized by gradient descent on reconstruction error at test time. An SAC policy conditioned on z selects actions. These changes distinguish the evaluated baseline from the published BATI procedure; in particular, test-time latent optimization is a property of this implementation. Because published BATI also models rewards, shared transition dynamics alone do not imply that it cannot distinguish nominal tasks.

Prompt-based Decision Transformer (prompt-DT) (Xu et al., 2022). Prompt-DT extends Decision Transformer (Chen et al., 2021) with trajectory prompts for few-shot policy generalization. For task T_i, a prompt τ_i^* of length K^* is sampled from demonstrations P_i, consisting of state-action-reward-to-go tuples $(s^*, a^*, \hat{G}^*)$. Appending the recent trajectory history τ_i from the offline dataset D_i gives $\tau_{input} = (\tau_i^*, \tau_i)$, a sequence of $3(K^* + K)$ tokens. The transformer autoregressively predicts actions for state tokens. We follow the original setup with $k = 20$ and use the same pretraining dataset as DPT, excluding query state-action pairs.

Adversarially Robust Decision Transformer (ARDT) (Tang et al., 2024). ARDT enhances the Decision Transformer by associating worst-case returns-to-go via minimax expectile regression with trajectories to improve robustness against adversarial perturbations. Specifically, the estimated Q values from expectile regression replaces the returns-to-go in vanilla DT during training.

Robust Adversarial Reinforcement Learning (RARL) (Pinto et al., 2017). RARL trains a protagonist and an adversary in a zero-sum Markov game: the protagonist maximizes return while the adversary applies destabilizing forces to minimize it. We use the framework from Reddi et al. (2024), with SAC policies for both agents.

Quantal Adversarial Reinforcement Learning (QARL) (Reddi et al., 2024). QARL adds entropy regularization to robust adversarial RL to model bounded rationality. Separate temperatures control the protagonist's and adversary's action entropies, yielding a Quantal Response Equilibrium (QRE). We use the framework from Reddi et al. (2024), with SAC policies for both agents.

PROTECTED (Liu et al., 2024). PROTECTED pretrains a finite set of non-dominated policies and uses adversarial-bandit feedback to select among them at deployment.

SeCBAD (Chen et al., 2022). SeCBAD jointly infers latent context and segment length, then conditions its policy on the inferred belief to adapt to piecewise-stationary environments.

C.2 Model Architecture and Hyperparameters

Transformer. DPT, MDPT, and prompt-DT use a causal GPT-2 architecture (Radford et al., 2019) with 4 attention layers and a single attention head per layer. ADAPT uses the same backbone with two independent projection heads for the task latent and adversary-centric latent. The embedding size is 32 in Dark Room and 256 in Meta-World and MuJoCo. Prompt-DT and ARDT embed transition components separately into a sequence; the remaining transformer-based models embed each (s, a, s', r) transition as a single token. We use AdamW with weight decay $1e-4$, learning rate $1e-3$, and batch size 128.

ADAPT-Specific Architecture and Objectives. ADAPT uses the same GPT-2 backbone described above with two independent projection heads, W_τ and W_ϕ, which project the query representation h_q to latent dimensions $d_z = 16$ for Dark Room and $d_z = 128$ for Meta-World and MuJoCo. The auxiliary loss weights in Equation 7 are set to $\lambda_1 = 0.1$ (task contrastive loss), $\lambda_2 = 1.0$ (dynamics SSL loss), and $\lambda_3 = 0.05$ (orthogonality constraint), with InfoNCE temperature $\kappa = 0.07$. The dynamics prediction head is a 3-layer MLP with 256 hidden units and ReLU activations. In continuous control environments (Meta-World, MuJoCo), the decoder D_ψ outputs a continuous action vector bounded by $\tanh$ scaling, optimized via mean squared error; for discrete action spaces (Dark Room), D_ψ outputs action logits optimized under standard categorical cross-entropy.

Multilayer perceptron (MLP). For all non-transformer agents, e.g., SAC, RARL, QARL, we directly list their shared hyper-parameters and architecture in Table 3.

Table 3: Non-transformer agent hyper-parameters (e.g, SAC, RARL, QARL)

Hyper-parameters	Values
No of hidden layers	3
No of hidden units per layer	256
activation function	ReLU
optimizer (actor and critic)	Adam
actor learning rate	$1 \cdot 10^{-4}$
critic learning rate	$3 \cdot 10^{-4}$
initial replay memory size	$3 \cdot 10^{3}$
max replay memory size	$1 \cdot 10^{6}$
warmup transitions	$5 \cdot 10^{3}$
batch size	256
target smoothing coefficient	$5 \cdot 10^{-3}$
initial temperature	$5 \cdot 10^{3}$
temperature learning rate	$3 \cdot 10^{-4}$

C.3 Computation and Implementation Efficiency

CORE collects the adversarial curriculum offline and distills it into an ICRL model following Algorithm 1. The collection cost comes from training experts and optimizing adversaries, but Phase I samples and independent Phase II initializations can be processed in parallel. The max-min iterations within each Phase II run remain sequential, because each adversary is optimized against the preceding policy. This distinguishes the offline cost of constructing the curriculum from deployment, where adaptation requires context-conditioned inference without parameter updates. All ADAPT pretraining experiments use a single NVIDIA RTX A5000 GPU (24GB). After data collection, ADAPT converges within 150 epochs ($<$ 1hour) in Dark Room, up to 500 episodes with early termination in Meta-World, and within 200 epochs ($<$ 2 hours) in MuJoCo.

D Environment Settings and Pretraining Dataset

D.1 Dark Room

Dark Room is a sparse-reward navigation task set in a discrete 10×10 grid. At the beginning of each episode, the agent is randomly placed in one of the grid cells, while the goal location is hidden and fixed at a random cell. The agent receives an observation of its current (x, y) position and selects from five discrete actions: move left, right, up, down, or stay in place. The episode horizon is $H = 100$ steps. The agent receives a reward of $r = 1$ only when it reaches the goal, and $r = 0$ otherwise. At test time, the agent always starts from the origin $(0, 0)$.

To construct the pretraining dataset for non-robust baselines, we generate $100,000$ trajectories using a uniform-random policy, evenly distributed across 100 different goal locations. For each, we sample query states uniformly and compute optimal actions using a heuristic that first aligns the agent's y-coordinate with the goal and then the x-coordinate. The dataset is partitioned into $80,000$ training examples (corresponding to 80 goal locations) and $20,000$ test examples (corresponding to 20 held-out goals used exclusively for testing).

For ADAPT, Phase I augments the pretraining tasks with adversarial perturbations. An adversarial action overrides the agent's intended action with probability $p \sim \mathcal{U}(0.0, 0.2)$. We sample an action-probability vector from a Dirichlet distribution with priors $\boldsymbol{\alpha}$ and select its most probable action. With probability $1 - p$, the agent executes its original action. Each perturbed task is a probabilistic MDP (Tessler et al., 2019), providing additional variation around the nominal goal. For ADAPT, Phase II of Algorithm 1 uses RARL (max-min optimization) (Pinto et al., 2017) to search for challenging perturbations and expand the curriculum.

The Dirichlet distribution used satisfies $\sum_{i=1}^{I} x_i = 1$ and $x_i \in [0, 1]$ for all i. The probability density function is given by:

$$f(x_1, ..., x_I; \alpha_1, ..., \alpha_I) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{i=1}^{I} x_i^{\alpha_i - 1}.$$

, where $\beta(\boldsymbol{\alpha})$ is the multivariate beta function, which can be expressed in terms of the gamma function

$$\beta(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{I} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{I} \alpha_i)}, \qquad \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_I).$$

These perturbed environments generate additional DPT-style trajectories, broadening pretraining experience for in-context generalization both with and without deployment disturbances.

D.2 Meta-World

We focus on the ML1 pick-place benchmark, where the agent must grasp an object and place it at a designated target location. Each task is defined by a 39-dimensional state space that includes the gripper's position and binary open/close state, the 3D pose of the object, and the coordinates of the target. The agent operates in a continuous action space, allowing it to adjust its end-effector position in three dimensions and control the gripper's open/close state to facilitate object manipulation. The environment provides shaped rewards to guide learning, including incentives for approaching the object, establishing a grasp, transporting the object, and releasing it at the target. The goal is to learn a policy that effectively sequences these skills to solve the overall manipulation task. For generalization evaluation, we train on 15 task configurations and test on 5 held-out tasks with novel object and target positions.

To construct the pretraining dataset, we collect historical trajectories from agents trained using Soft Actor-Critic (SAC). For each task, SAC is trained until convergence, and we sample from the resulting trajectories to form offline datasets. The built-in deterministic policy provides expert action targets. To simulate real-world deployment challenges, we introduce robustness through adversarial perturbations to the observation space. These perturbations target only the perceptually estimated components of the observation, namely, object pose, end-effector position, and goal location, while



(a) Cartpole (b) Cheetah (c) Hopper (d) Quadruped (e) Reacher (f) Walker

Figure 5: Illustrations of the MuJoCo environments.

avoiding internal robot states such as joint angles or velocities, which are typically less affected by real-world noise. This design mimics sensor uncertainty (e.g., from depth cameras or visual drift) without destabilizing control signals that are precise on physical hardware. Following Phase I of Algorithm 1, Phase II uses adversarial RL (Pinto et al., 2017) to search for challenging perturbations and retain the corresponding expert policies. Combining these curricula across nominal training tasks provides ADAPT with experience in task variation and disturbed execution.

D.3 MuJoCo

We use MuJoCo environments (Todorov et al., 2012) from the DeepMind Control Suite (Tunyasuvunakool et al., 2020), modified to support adversarial forces for RARL, QARL, and pretraining ADAPT. The adversary's action space differs from the protagonist's and applies forces to selected body parts. Environment-specific force bounds and action spaces are listed in Table 4.

Table 4: MuJoCo Environment-specific parameters for adversarial training and (protagonist) agent's standard observation/action space

Environment	Adversary max force	Adversary action space	Observation space	Action space
Cartpole	0.005	2D forces on pole (1)	5	1
Cheetah	1.0	2D force on feet & torso (6)	17	6
Hopper	1.0	2D force on feet & torso (4)	15	4
Quadruped	10	3D force on torso (1)	78	12
Reacher	0.1	2D force on arm (2)	6	2
Walker	1.0	2D force on feet (4)	24	6

We conduct experiments in 6 environments (Figure 5), with observation and action spaces listed in Table 4. The discount factor is 0.9 and the horizon is 200 for all environments.

Each environment is associated with one or more problems, defined as instances of the model with specific Markov Decision Process (MDP) structures. For example, in the CartPole environment, we define two problems: **swingup**, where the pole starts pointing downward, and **balance**, where the pole begins near the upright position. The goal in both problems is to manipulate the forces applied to a cart at the base in order to either swing up or balance an unactuated pole. In the Reacher environment, a two-link planar arm must reach a randomized target location, with a reward of 1 when the end effector reaches the target sphere. Two problems are defined here: in the **easy** problem, the target sphere is larger than in the **hard** problem.

In the Hopper environment, a planar one-legged hopper is initialized in a random configuration and is rewarded based on torso height and forward velocity. The remaining three environments: Walker, Cheetah, and Quadruped focus on maximizing forward velocity. In the Cheetah environment, the reward is linearly proportional to the forward velocity, capped at a maximum of $10m/s$. The Walker environment includes two tasks: **walk** and **run**, which differ in their velocity requirements and include components to encourage upright posture and minimal lateral movement. For standard in-context reinforcement learning (ICRL) methods, such as DPT (Lee et al., 2023), we use m pretraining tasks $\{\tau^i\}_{i=1}^m \subset \mathcal{M}$ with varying internal conditions. For each task τ^i, we construct a pretraining dataset using 6000 historical trajectories collected from SAC agents trained to convergence. We follow the SAC training approach in Reddi et al. (2024), training SAC policies under different variations, which are treated as distinct tasks, and then further solve the max-min optimization upon the trained SAC policies for boundary expansion in Algorithm 1.

Table 5 summarizes the CORE pretraining curriculum hyperparameters for three benchmarks as well as the number of trajectories per task/goal.

Table 5: **CORE pretraining curriculum hyperparameters across benchmark domains.**

Benchmark	N_{I}	N_{II}	T	Total Trajectories per Task/Goal
Dark Room	5	1	5	1,000
Meta-World	10	5	2	6,000
MuJoCo	10	5	10	6,000

E MuJoCo Results

We evaluate our methods on a diverse set of continuous control tasks from the DeepMind Control Suite (Tunyasuvunakool et al., 2020), spanning 6 environments and 11 tasks (details in Appendix D). Pretraining datasets are constructed from historical trajectories generated by SAC agents. For adversarial data generation, each adversary is modeled by a distinct neural network. The MuJoCo adversary is based on active external forces applied within a zero-sum Markov Game framework. The adversary's policy is parameterized to apply forces that destabilize the agent. The specific adversary action space and maximum force bounds are determined by domain knowledge and vary per environment. SAC is trained to convergence under adversarial disturbances, and the resulting trajectories are used to build variation environments for pretraining tasks. Additional dataset details are provided in Appendix D. We assess robustness from two complementary perspectives: (a) adversarial disturbances affecting the agent and (b) environmental variations, such as changes in mass and friction.

RARL and QARL both use SAC agents, making SAC the corresponding baseline without adversarial training. Our adversarial setup follows Reddi et al. (2024), with parameters in Table 4. We evaluate robustness to adversarial forces and generalization to changes in physical parameters.

Robustness to Agent Disturbance. We evaluate learned policies against an adversary trained to minimize the agent's return, with the evaluated policy parameters fixed.

Robustness to Environmental Change. We evaluate how well policies generalize to environments with variations in physical parameters such as mass and friction. Generalization is assessed by deploying policies in environments with shifted dynamics coefficients.

We group MuJoCo tasks into cartpole, reacher, and locomotion (quadruped, cheetah, walker, and hopper). The box plots evaluate policies at the nominal physical parameters used to train the standard RL baselines, both without and with adversarial forces. The heatmaps separately evaluate generalization across physical parameters, including values outside the ICRL pretraining distribution.

Figures 6–14 report the box plots and heatmaps; PDT denotes prompt-DT (Xu et al., 2022), and MDPT denotes mix-DPT (Wang et al., 2024). ADAPT improves robustness to adversarial forces compared with DPT. Training with adversarial variations can also improve undisturbed performance, consistent with findings in robust RL (Pinto et al., 2017; Dong et al., 2025b). The gains are particularly clear in high-dimensional control, including Quadruped-Run (Figure 8).

The heatmaps in Figure 9–14 evaluate generalization to unseen physical parameters. Each 8×8 grid varies two environment parameters, while ICRL pretraining covers only the central 4×4 grid. Thus, cells outside this region test unseen nominal task configurations. ADAPT shows strong performance across these shifts, with clear gains in Walker-Run and Cheetah-Run (Figures 12b and 13a).

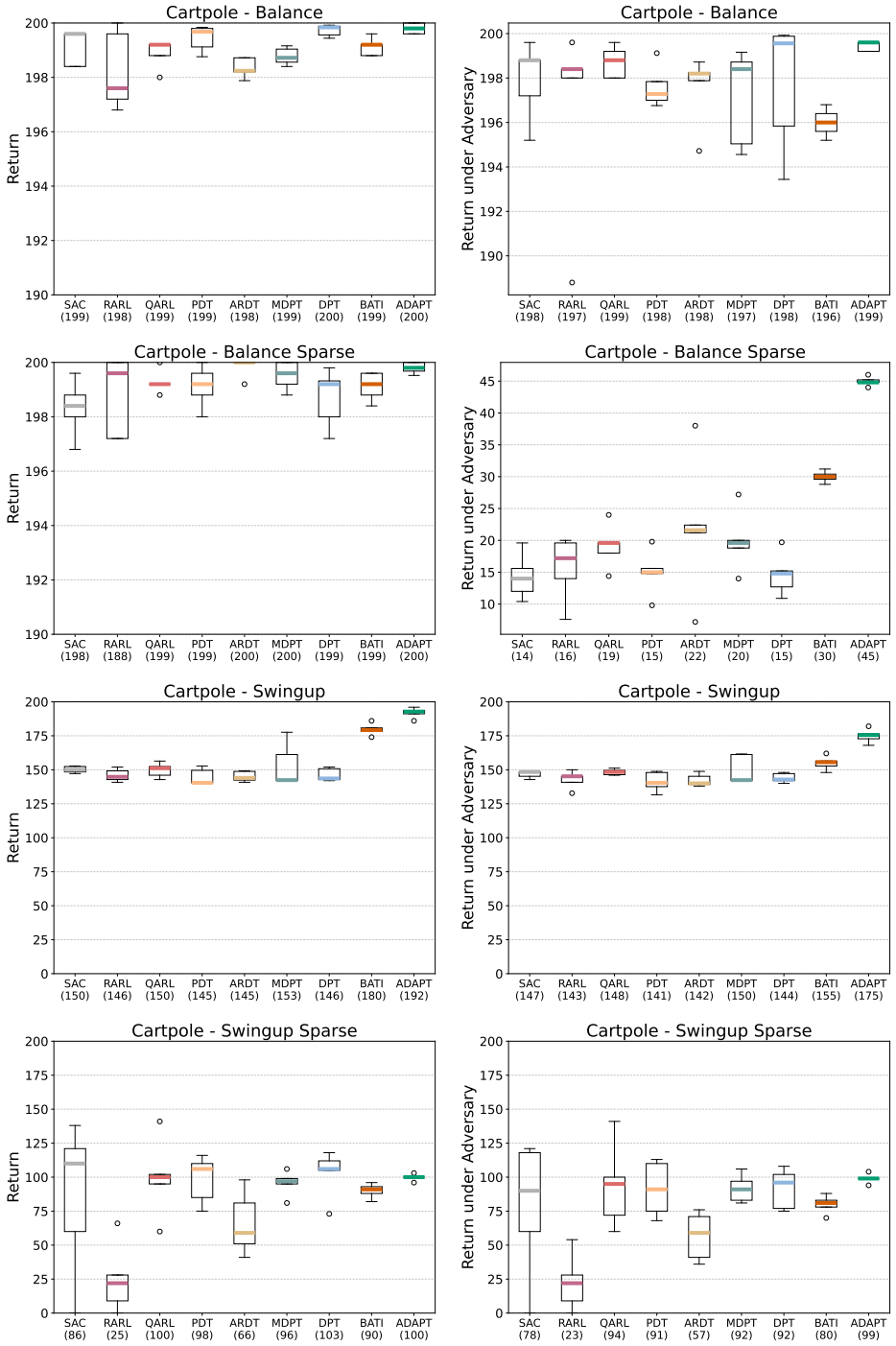


Figure 6: Performance and robustness on Cartpole problems (see the title of each boxplot), evaluated at the end of training without an adversary (left column) and against an adversary (right column). The number next to each algorithm is the average performance across 10 seeds.

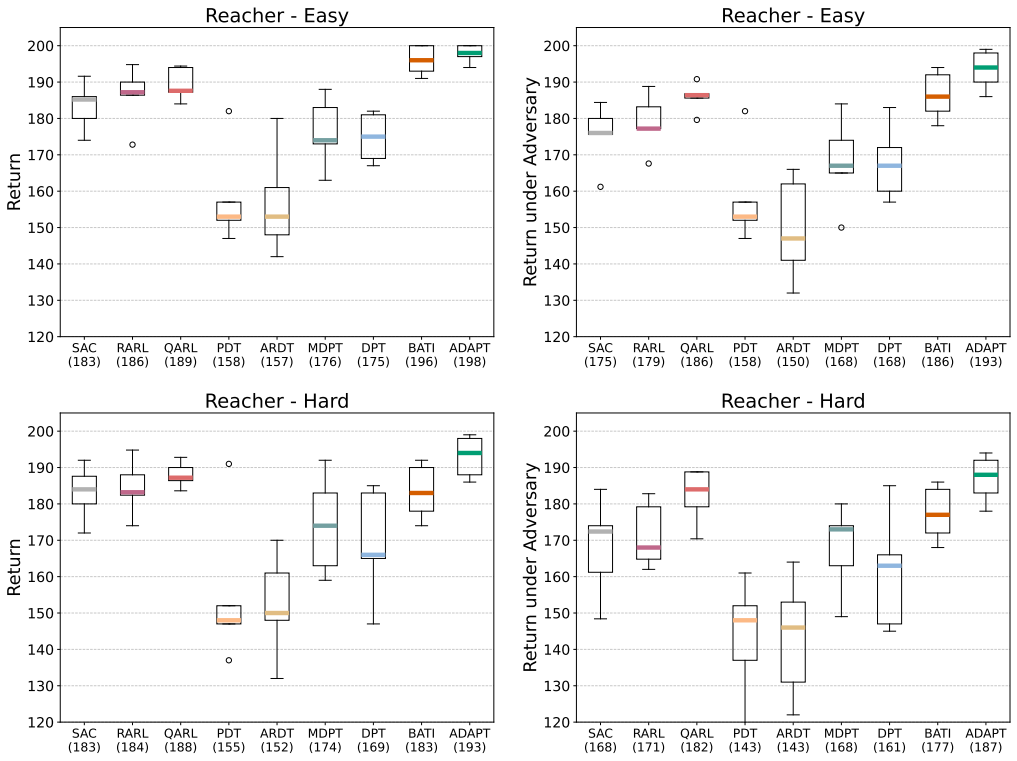


Figure 7: Performance and robustness on Reacher problems (see the title of each boxplot), evaluated at the end of training without an adversary (left column) and against an adversary (right column). The number next to each algorithm is the average performance across 10 seeds.

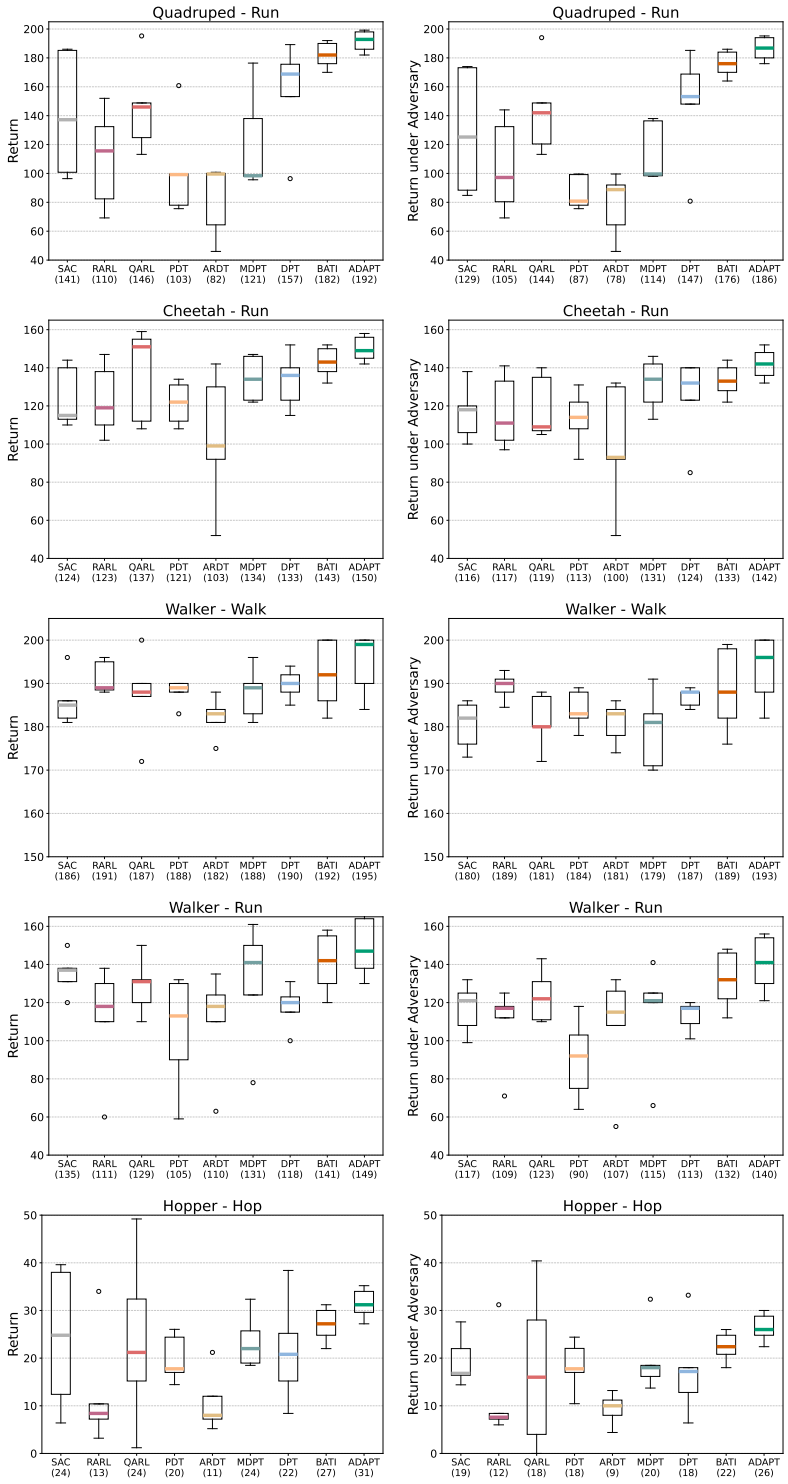
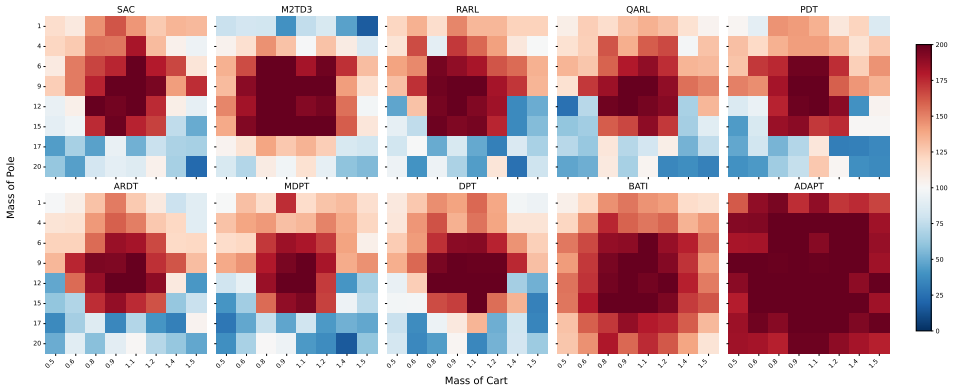
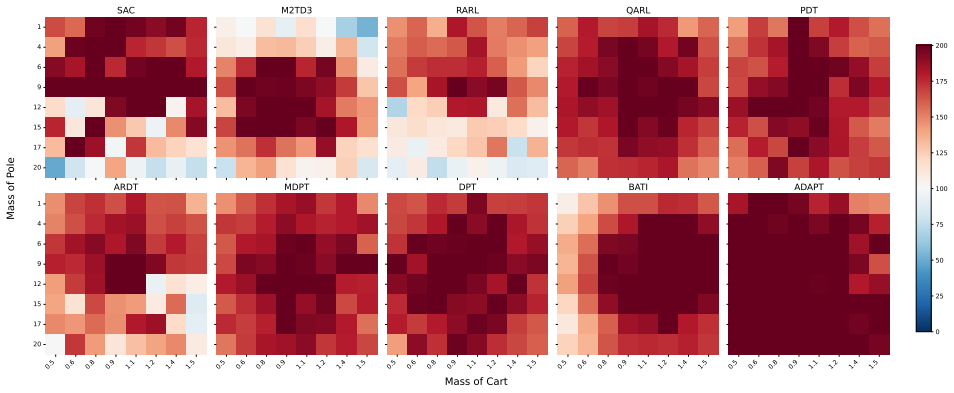


Figure 8: Performance and robustness on locomotion problems, i.e., Quadruped, Cheetah, Walker, and Hopper (see the title of each boxplot), evaluated at the end of training without an adversary (left column) and against an adversary (right column). The number next to each algorithm is the average performance across 10 seeds.

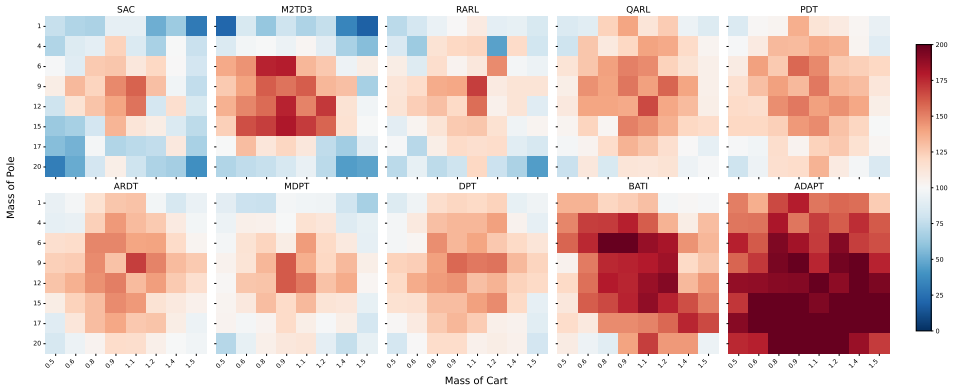


(a) Cartpole-Balance

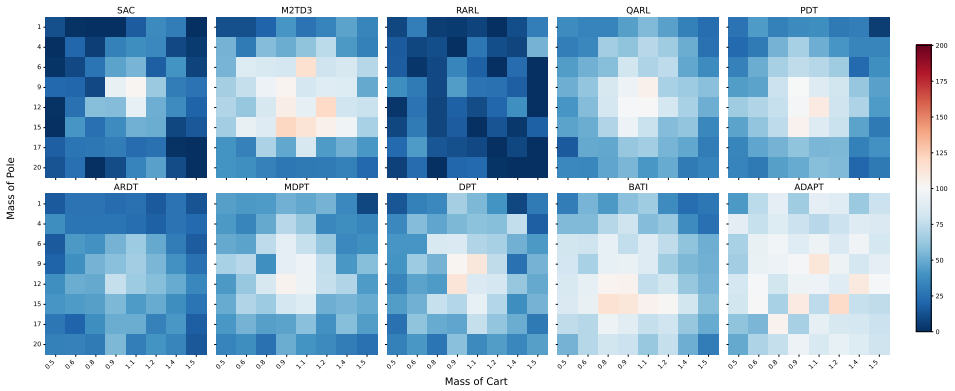


(b) Cartpole-Balance-Sparse

Figure 9: Robustness analysis on Cartpole Balance (see the title of each heatmap set). Heatmaps show mean returns across 10 seeds for varying environment properties (x-y axes). The central 4×4 grid is covered by ICRL pretraining; the remaining cells evaluate unseen parameter configurations.

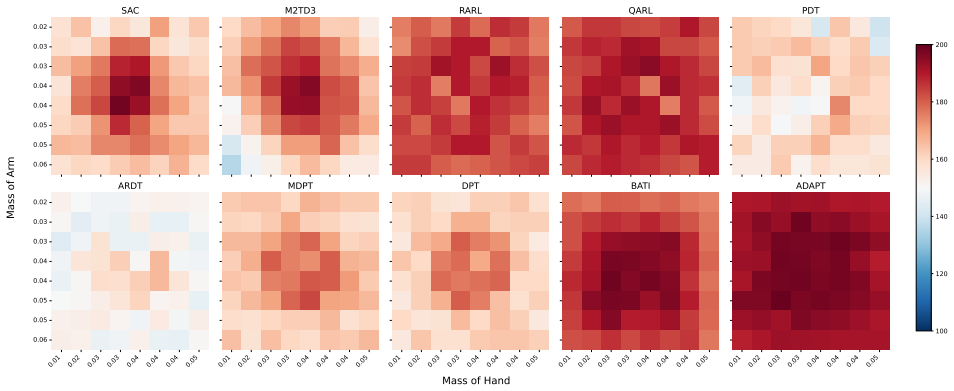


(a) Cartpole-Swingup

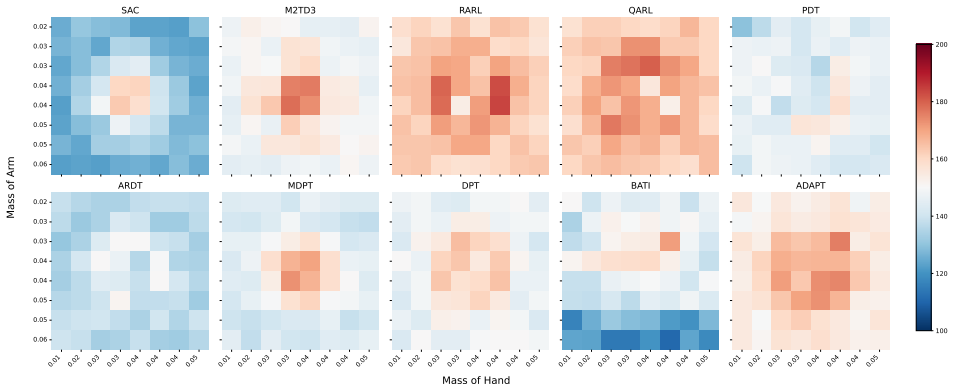


(b) Cartpole-Swingup-Sparse

Figure 10: Robustness analysis on Cartpole Swingup (see the title of each heatmap set). Heatmaps show mean returns across 10 seeds for varying environment properties (x-y axes). The central 4×4 grid is covered by ICRL pretraining; the remaining cells evaluate unseen parameter configurations.

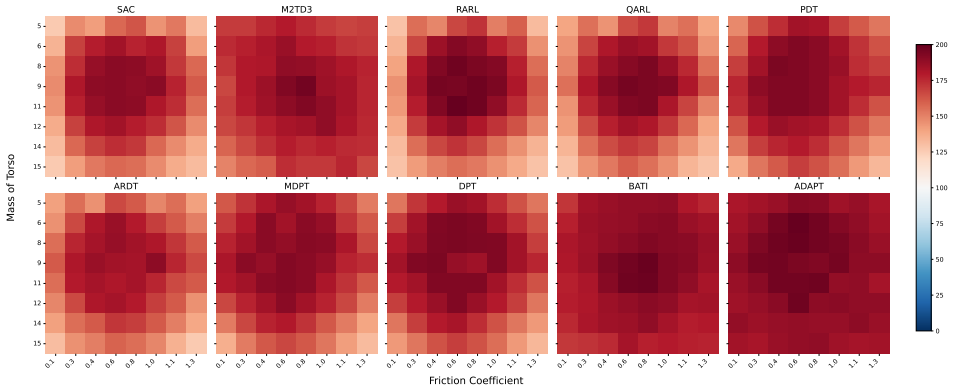


(a) Reacher-Easy

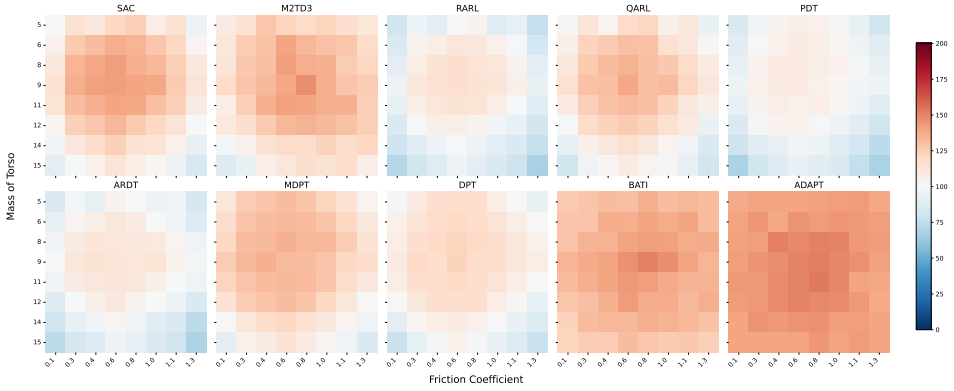


(b) Reacher-Hard

Figure 11: Robustness analysis on Reacher (see the title of each heatmap set). Heatmaps show mean returns across 10 seeds for varying environment properties (x-y axes). The central 4×4 grid is covered by ICRL pretraining; the remaining cells evaluate unseen parameter configurations.

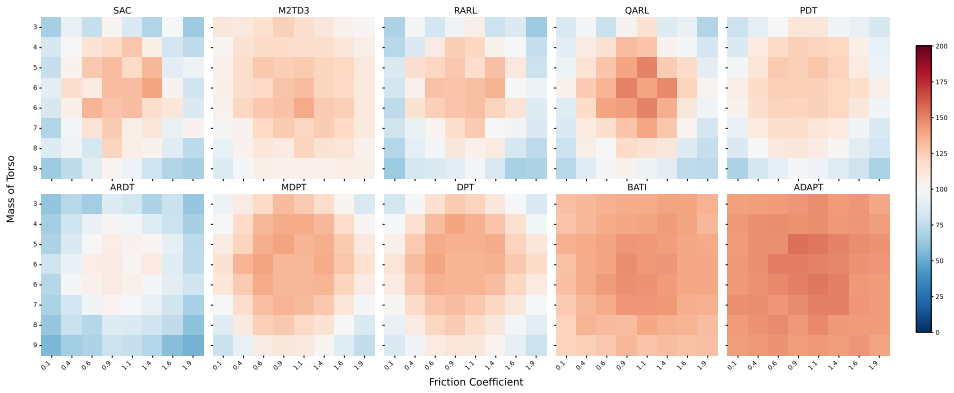


(a) Walker-Walk

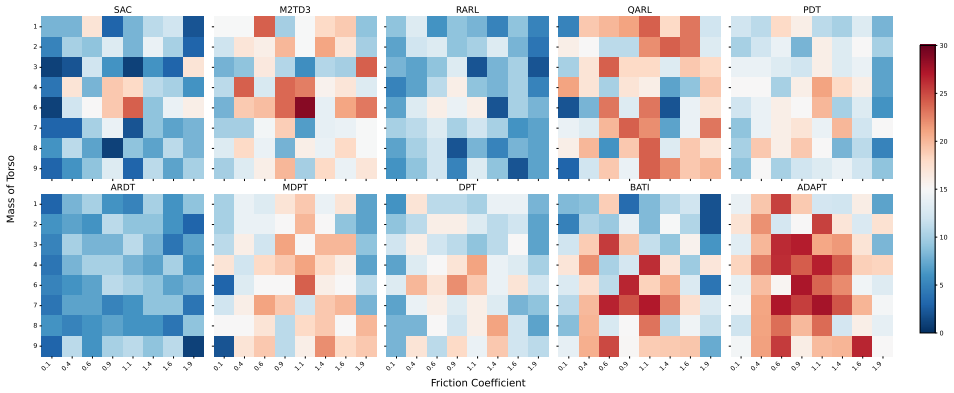


(b) Walker-Run

Figure 12: Robustness analysis on locomotion problems (Walker). Heatmaps show mean returns across 10 seeds for varying environment properties (x-y axes). The central 4×4 grid is covered by ICRL pretraining; the remaining cells evaluate unseen parameter configurations.

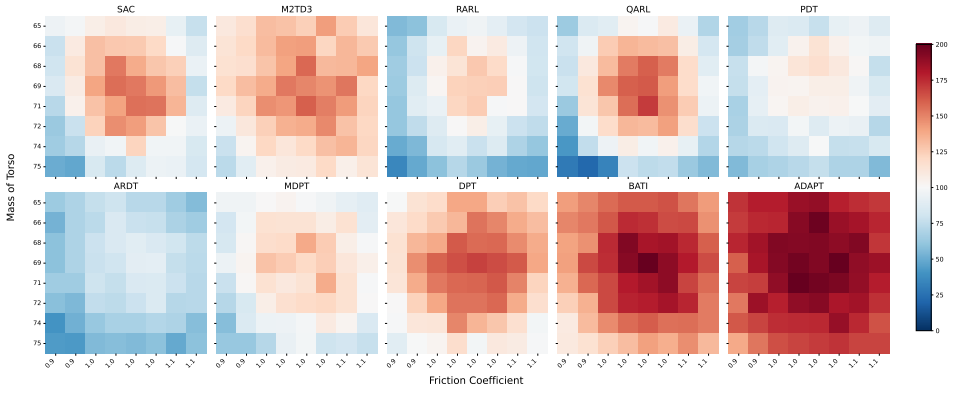


(a) Cheetah-Run



(b) Hopper-Hop

Figure 13: Robustness analysis on locomotion problems (Cheetah and Hopper). Heatmaps show mean returns across 10 seeds for varying environment properties (x-y axes). The central 4×4 grid is covered by ICRL pretraining; the remaining cells evaluate unseen parameter configurations.



(a) Quadruped-Run

Figure 14: Robustness analysis on locomotion problems (Quadruped). Heatmaps show mean returns across 10 seeds for varying environment properties (x-y axes). The central 4×4 grid is covered by ICRL pretraining; the remaining cells evaluate unseen parameter configurations.

F Ablation Studies

The main experiments evaluate robust control and generalization to unseen tasks. Here, we examine the contributions of the mechanisms developed in Sections 3 and 4. We first hold the ADAPT architecture fixed to study the CORE curriculum and robust context collection. We then examine the architecture and its representation objectives, followed by linear probes that assess the task and adversary information retained in the learned latents.

F.1 Efficacy of the CORE Pretraining Pipeline

We first evaluate CORE's two-phase curriculum and robust context collection while keeping the downstream ADAPT architecture fixed (Section 3, Algorithm 1). This separates the effect of the pretraining experience and initial behavior from the representation design. In Table 6, adding boundary-focused experts from Phase II to the uniform coverage of Phase I raises Episode 40 return from 6.5 ± 0.5 to 13.1 ± 0.5. Allocating the full Phase I + II budget to Phase I yields 7.4 ± 0.6 at Episode 40, confirming that the gain from adversarial search persists under the same budget. Robust initialization has its largest relative effect early: Episode 5 return rises from 3.8 ± 0.6 to 6.5 ± 0.4, while Episode 40 return further improves to 13.9 ± 0.2. The two mechanisms thus make complementary contributions: the curriculum supplies the experience from which adaptation is learned, and the initial behavioral policy determines the context available when deployment begins.

Table 6: **CORE curriculum and robust-initialization ablation with the ADAPT architecture fixed.** Evaluated on held-out Meta-World ML1 pick-place configurations under disturbance. Values report average undiscounted cumulative episodic return across 10 evaluation seeds ($\pm$ standard error of the mean).

Setup	Episode 5	Episode 10	Episode 20	Episode 40
Phase I	1.6 ± 0.4	2.4 ± 0.6	2.8 ± 0.5	6.5 ± 0.5
Phase I (full budget)	1.9 ± 0.5	2.9 ± 0.5	3.6 ± 0.6	7.4 ± 0.6
Phase I + II	3.8 ± 0.6	7.4 ± 0.5	10.2 ± 0.5	13.1 ± 0.5
Phase I + II (Robust Initialization)	$\mathbf{6.5 \pm 0.4}$	$\mathbf{9.7 \pm 0.3}$	$\mathbf{12.6 \pm 0.3}$	$\mathbf{13.9 \pm 0.2}$

F.2 Architectural Contributions of ADAPT

Having examined the curriculum and initial context, we turn to the representation objectives introduced in Section 4. Table 2 starts with a single-stream baseline, replaces it with a dual-stream architecture trained with task anchoring ($\mathcal{L}_{task}$), and then adds the self-supervised dynamics loss ($\mathcal{L}_{\text{SSL}}$) and orthogonality constraint ($\mathcal{L}_{\text{ortho}}$). Returns improve with each addition across Meta-World and three MuJoCo tasks. This progression supports the motivation for assigning complementary roles to the task and adversary representations: the gains are not explained by adversarial pretraining alone, but also depend on how the resulting experience is represented.

F.3 Task and Adversary Information in the Learned Latents

The return ablations show that the representation objectives improve decision-making, but do not directly reveal which information the latents encode. We therefore train linear probes on the frozen task and adversary latents (z_τ and z_ϕ), as well as the single-stream feature (h). Using 5000 Meta-World contexts (80/20 train/test split, 5 seeds), probes are trained to classify (i) *task identity* across the 15 training object-target configurations, and (ii) *adversary identity* across 8 distinct perturbation targets (e.g., object pose, end-effector position, and goal location). For task-identity prediction, Table 7 reports 91% accuracy from the task latent, compared with 35% from the adversary latent and 58% from the single-stream feature. Conversely, adversary-identity prediction is more accurate from the adversary latent (85%) than from the task latent (26%) or the single-stream feature (51%). These comparisons hold the prediction target fixed and support the intended allocation of task and disturbance information across the two latents, complementing the return improvements in the component ablations.

Table 7: Linear probing accuracy (%) on Meta-World frozen representations. Within each row, all representations predict the same target; the designated latent gives the highest accuracy.

Probe Target	**From z_τ (Task Latent)**	**From z_ϕ (Adversary Latent)**	**From Single-Stream h**
Task Identity	**91%**	35%	58%
Adversary Identity	26%	**85%**	51%

G Proofs and Extensions of the Generalization Guarantees

G.1 Full Statement of the Performance Guarantee

The main theorem isolates the advantage of tailored control. We give its statistical assumptions and explicit error terms here, using a common setup for a fixed nominal task and for fresh nominal tasks. Consider horizon-H MDPs with rewards in $[0, 1]$, indexed by e: for a fixed nominal task, $e = \psi$ is the disturbance; for the task-family setting, $e = (\tau, \psi)$. This index remains fixed during context collection and evaluation. Let π_e^E denote a curriculum expert, which need not be optimal, and let Q be the pretraining population of $(e, X, A^\star)$, where $X = (\xi, s, h)$ consists of context, query state, and time. Labels satisfy $A^\star \sim \pi_e^E(\cdot \mid s, h)$, conditionally independently of context given (e, s, h). The population action predictor is

$$q_N(\cdot \mid X) = \mathbb{E}_Q[\pi_e^E(\cdot \mid s, h) \mid X], \qquad \mathcal{I}_N = I_Q(A^\star; e \mid X). \tag{11}$$

The residual uncertainty $\mathcal{I}_N$ concerns the expert action after context and query have been observed. It can be small without identifying every aspect of the environment. This supervised ICRL formulation (Lee et al., 2023; Lin et al., 2023) uses a posterior expert predictor; it does not require explicit storage or selection of experts at deployment.

Write $\bar{J}_\theta(e) = \mathbb{E}_{\xi \sim B_e^N} J(T_\theta(\xi), e)$ for post-context return under the actual collector's context law B_e^N, with parameters and context fixed during the evaluation episode. Set $f_\theta(e) = J^*(e) - \bar{J}_\theta(e)$. The deployment law R_θ of (e, ξ, s, h) draws e from the same population as Q, collects context under B_e^N, and samples time uniformly and state from the learner's corresponding occupancy. The joint coverage condition $R_\theta \le CQ_{e,X}$ accounts for both query-state coverage and differences between pretraining and deployment context laws. We impose it for candidate policies as specified in Appendix G.2; context informativeness is measured separately by $\mathcal{I}_N$.

Theorem G.1 (Detailed form of Theorem 5.1)**.** *Assume joint coverage as above. Pretraining uses M independent, identically distributed task collections, with arbitrary dependence among examples within each collection and equal weight per collection. Let Q be the marginal law of a uniformly sampled example from a random collection. Suppose a fixed, data-independent class of K predictors contains q_N, and the learned predictor's empirical action loss is at most that of q_N plus ζ_M. Then, with probability at least $1 - \delta$,*

$$\mathbb{E}_e f_{\widehat{\theta}}(e) \le \varepsilon_{N,M} := \epsilon_E + H^2 \left(\sqrt{\frac{C\mathcal{I}_N}{2}} + \sqrt{C\zeta_M + \frac{2C \log(K/\delta)}{M}} \right), \tag{12}$$

where $\epsilon_E = \mathbb{E}_e[J^*(e) - J(\pi_e^E, e)]$. For $e = \psi$, suppose additionally that Ψ is compact, its curriculum distribution satisfies $P_\Psi(B_r(\psi)) \ge cr^d$ for every ψ and $0 < r \le r_0$, and $f_{\widehat{\theta}}$ is L-Lipschitz. Define

$$\mathcal{B}_d(u) = \min \left\{ H, \inf_{0 < r \le r_0} \left[\frac{u}{cr^d} + Lr \right] \right\}. \tag{13}$$

On the same event, the worst-case gain satisfies equation 9.

The proof appears in Appendix G.5. The sample count M is the number of independent collections, rather than correlated transitions. Lemma G.4 allows unrestricted dependence within a collection; Lemma G.5 and equation 34 give approximation and covering variants of the finite-class condition. The expert, information, and prediction terms are separated by Proposition G.3.

For Corollary G.10, μ is a population prior over nominal tasks, rather than the empirical distribution of the finite training set. Both $V^{\mathrm{rob}}(\tau)$ and $\Delta_{expert}(\tau)$ are defined taskwise by equation 8. The

corollary uses common neighborhood-mass and regret-Lipschitz constants within tasks. Its collector is an available policy or protocol satisfying joint coverage; the comparator $V^{\mathrm{rob}}(\tau)$ does not assume that a task-specific robust controller is available before an unseen task is identified.

G.2 Statistical Setup and Scope

We consider finite-horizon, fully observed MDPs with common state and action spaces and rewards in $[0, 1]$. Time-dependent policies are allowed. The environment index e fixes the initial-state, reward, and transition laws throughout the collection and evaluation episodes. Random disturbances governed by these fixed laws are allowed; an adversary that changes the latent law in response to the deployment history requires a different analysis. All random variables take values in standard Borel spaces so that regular conditional distributions exist. Returns include initial-state, transition, and action randomness.

The population training law factors as

$$Q(de, d\xi, ds, dh, da) = P(de)Q_e^{\mathrm{pre}}(d\xi)\frac{1}{H}\nu_{e,h}(ds \mid \xi)\pi_e^E(da \mid s, h). \tag{14}$$

In particular, the expert label is conditionally independent of context given (e, s, h). The query distribution may depend on both e and ξ, which is why the posterior in equation 11 conditions on the query. A finite collection of dependent examples is permitted provided its uniformly selected example has this law. Any learned expert or curriculum generator used to specify Q is fixed, or conditioned on, before invoking independence of the pretraining collections.

For a frozen predictor T_θ and collected context ξ, let $d_{e,h}^{\theta,\xi}$ be its evaluation occupancy. The deployment law is

$$R_\theta(de, d\xi, ds, dh) = P(de)B_e^N(d\xi)\frac{1}{H}d_{e,h}^{\theta,\xi}(ds).$$

The measure inequality $R_\theta \le CQ_{e,X}$ means that its Radon–Nikodym derivative is bounded by C almost surely. When collection laws match, conditional query-state coverage suffices. When they differ, state coverage alone does not suffice. We assume this domination for every candidate predictor, or simultaneously on the event to which the learning theorem is applied. Thus selecting a predictor from data does not invalidate the change of measure. The collector is an actual available policy or protocol, not an oracle inferred from the comparator in the return bound.

We use $\mathrm{TV}(p, q) = \frac{1}{2}\int |p - q|$ and $h^2(p, q) = 1 - \int \sqrt{pq}$. The information-based theorem applies to finite actions or action distributions with well-defined conditional KL divergences. It is informative when $\mathcal{I}_N < \infty$. In continuous spaces, deterministic experts with different actions can have infinite conditional mutual information; Proposition G.8 supplies a value-based alternative. All statements are about expected post-context returns. They do not guarantee improvement for every realized context, or describe the entire process of updating context within an evaluation episode.

G.3 Posterior Prediction and the Average Return Bound

Lemma G.2 (Posterior prediction and residual uncertainty)**.** *Under equation 14, the conditional expert-label law is $q_N = Q(A^\star \mid X)$ as in equation 11, and*

$$\mathbb{E}_Q \mathrm{KL}(\pi_e^E(\cdot \mid s, h) \| q_N(\cdot \mid X)) = I_Q(A^\star; e \mid X). \tag{15}$$

Whenever the cross-entropies are finite, the population action loss satisfies

$$\mathcal{L}_Q(\theta) - \mathcal{H}_Q(A^\star \mid X) = \mathbb{E}_{Q_X} \mathrm{KL}(q_N(\cdot \mid X) \| T_\theta(\cdot \mid X)). \tag{16}$$

Thus q_N minimizes the unrestricted population action loss, uniquely as a conditional distribution up to Q_X-null sets, not necessarily as a parameter vector.

Proof. Conditional expectation in equation 14 gives the posterior mixture defining q_N. Conditional mutual information is the expectation of the KL divergence between $Q(A^\star \mid e, X)$ and $Q(A^\star \mid X)$. The first distribution equals $\pi_e^E(\cdot \mid s, h)$, establishing equation 15. Adding and subtracting $\log q_N(A^\star \mid X)$ inside the action loss proves equation 16. Nonnegativity of conditional KL establishes the minimization claim. No trajectory-distribution KL is needed. □

Proposition G.3 (Average performance with imperfect experts)**.** *For any frozen predictor satisfying joint coverage, let* $\eta_\theta = \mathbb{E}_{R_\theta} \mathrm{TV}(q_N(\cdot \mid X), T_\theta(\cdot \mid X))$*. Then*

$$\mathbb{E}_e f_\theta(e) \le \epsilon_E + H^2 \left(\sqrt{C\mathcal{I}_N/2} + \eta_\theta \right). \tag{17}$$

If a bound g *on the right-hand side of equation 16 is available, one may take* $\eta_\theta \le \sqrt{Cg/2}$ *instead.*

Proof. Fix e and ξ. The finite-horizon performance-difference identity, using the expert's value functions and the learner's occupancy, gives

$$J(\pi_e^E, e) - J(T_\theta(\xi), e) = \sum_{h=1}^{H} \mathbb{E}_{s \sim d_{e,h}^{\theta,\xi}} \left[\mathbb{E}_{a \sim \pi_e^E} Q_{e,h}^{\pi_e^E}(s,a) - \mathbb{E}_{a \sim T_\theta} Q_{e,h}^{\pi_e^E}(s,a) \right].$$

Since $Q_{e,h}^{\pi_e^E}$ takes values in $[0, H-h+1]$, each difference is at most $H \,\mathrm{TV}(\pi_e^E, T_\theta)$. Add $J^*(e) - J(\pi_e^E, e)$ and average over e, context, and uniform time. The triangle inequality gives

$$\mathbb{E}_e f_\theta(e) \le \epsilon_E + H^2 \left[\mathbb{E}_{R_\theta} \mathrm{TV}(\pi_e^E, q_N) + \eta_\theta \right].$$

Pinsker's inequality, Jensen's inequality, joint domination, and equation 15 yield

$$\mathbb{E}_{R_\theta} \mathrm{TV}(\pi_e^E, q_N) \le \sqrt{\tfrac{1}{2} \mathbb{E}_{R_\theta} \mathrm{KL}(\pi_e^E \| q_N)} \le \sqrt{C\mathcal{I}_N/2}.$$

The same argument bounds the prediction term by $\sqrt{Cg/2}$. The expert need not be optimal; its return error is precisely the separate term ϵ_E. □

G.4 Learning from Independent Nominal-Task Blocks

A pretraining curriculum can contain many dependent examples from the same nominal task. We therefore count independent *task blocks*, rather than treating all context–query pairs as independent task draws. Let

$$\mathcal{B}_i = \{(X_{ij}, A_{ij}^\star)\}_{j=1}^{n_i}, \qquad i = 1, \dots, M,$$

be independent and identically distributed random blocks, with $1 \le n_i < \infty$ almost surely. Here $X = (\xi, s, h)$ contains the context, query state, and decision time. Dependence within a block is unrestricted: examples may share a nominal task, trajectories, experts, or an adaptively generated adversarial curriculum. Define Q as the distribution obtained by drawing one population block and then choosing one of its examples uniformly. In particular, for every integrable function f,

$$\mathbb{E}_{(X,A^\star) \sim Q} f(X, A^\star) = \mathbb{E}_{\mathcal{B}} \left[\frac{1}{n(\mathcal{B})} \sum_{j=1}^{n(\mathcal{B})} f(X_j, A_j^\star) \right]. \tag{18}$$

The empirical objective gives equal weight to each task block:

$$\widehat{\mathcal{L}}(p) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{n_i} \sum_{j=1}^{n_i} -\log p(A_{ij}^\star \mid X_{ij}). \tag{19}$$

When block sizes are equal, this is the usual average over all examples. Pooling unequal block sizes without normalization changes the weighting of nominal tasks and requires a corresponding sampling analysis.

Let $q(\cdot \mid x) = Q(A^\star \in \cdot \mid X = x)$ be the population conditional action distribution. For discrete actions, all expressions below use probability mass functions. More generally, assume that q and the candidate policies admit measurable conditional densities with respect to a common dominating measure λ. We use the squared Hellinger convention

$$h^2(p, q \mid x) := 1 - \int \sqrt{p(a \mid x) q(a \mid x)} \, \mathrm{d}\lambda(a), \tag{20}$$

$$h_Q^2(p, q) := \mathbb{E}_{X \sim Q_X} h^2(p, q \mid X). \tag{21}$$

For this convention, $\mathrm{TV}^2(p(\cdot \mid x), q(\cdot \mid x)) \le 2h^2(p, q \mid x)$.

Lemma G.4 (Conditional likelihood learning from task blocks)**.** *Let $\Pi = \{p_1, \dots, p_K\}$ be a finite, data-independent class of conditional action distributions. Under the block sampling scheme above, with probability at least $1 - \delta$, every $p \in \Pi$ satisfies*

$$h_Q^2(q, p) \le \frac{\widehat{\mathcal{L}}(p) - \widehat{\mathcal{L}}(q)}{2} + \frac{\log(K/\delta)}{M}. \tag{22}$$

Consequently, any data-dependent $\widehat{p} \in \Pi$ satisfying $\widehat{\mathcal{L}}(\widehat{p}) \le \widehat{\mathcal{L}}(q) + \zeta$, for $\zeta \ge 0$, obeys

$$h_Q^2(q, \widehat{p}) \le \frac{\zeta}{2} + \frac{\log(K/\delta)}{M}. \tag{23}$$

No membership assumption $q \in \Pi$ is required for this implication. The slack ζ is measured relative to the true conditional distribution q, and is not generally an optimization error alone. If $q \in \Pi$, approximate empirical risk minimization over Π with optimization error at most ζ is sufficient to meet the stated comparison.

If a deployment query distribution R_X satisfies $R_X \ll Q_X$ and $\mathrm{d}R_X/\mathrm{d}Q_X \le C$, then, on the same event,

$$\mathbb{E}_{X \sim R_X} \mathrm{TV}(q(\cdot \mid X), \widehat{p}(\cdot \mid X)) \le \sqrt{C\zeta + \frac{2C \log(K/\delta)}{M}}. \tag{24}$$

Proof. For a fixed candidate p, define the block log-likelihood ratio

$$\Delta_i(p) = \frac{1}{n_i} \sum_{j=1}^{n_i} \log \frac{q(A_{ij}^\star \mid X_{ij})}{p(A_{ij}^\star \mid X_{ij})}.$$

The reference density $q(A^\star \mid X)$ is positive and finite almost surely under Q, after choosing the usual density versions. A zero candidate density on a sampled action gives a log-likelihood ratio of $+\infty$ and an exponential of zero. Thus the following argument also covers candidates with zero likelihood. Empirical differences are interpreted through these log-likelihood ratios; no finite population entropy assumption is needed for this lemma.

By convexity of the exponential, pointwise within each finite block,

$$\exp\left(-\frac{\Delta_i(p)}{2}\right) \le \frac{1}{n_i} \sum_{j=1}^{n_i} \sqrt{\frac{p(A_{ij}^\star \mid X_{ij})}{q(A_{ij}^\star \mid X_{ij})}}. \tag{25}$$

Taking expectations and using equation 18 yields

$$\begin{aligned} \mathbb{E} \exp\left(-\frac{\Delta_i(p)}{2}\right) &\le \mathbb{E}_{(X, A^\star) \sim Q} \sqrt{\frac{p(A^\star \mid X)}{q(A^\star \mid X)}} \\ &= \mathbb{E}_{X \sim Q_X} \int \sqrt{q(a \mid X) p(a \mid X)}\, \mathrm{d}\lambda(a) \\ &= 1 - h_Q^2(q, p). \end{aligned} \tag{26}$$

This step uses no independence among examples in the same block. Independence across blocks gives

$$\mathbb{E} \exp\left[-\frac{1}{2} \sum_{i=1}^{M} \Delta_i(p)\right] \le [1 - h_Q^2(q, p)]^M \le \exp[-M h_Q^2(q, p)]. \tag{27}$$

Since $M^{-1} \sum_i \Delta_i(p) = \widehat{\mathcal{L}}(p) - \widehat{\mathcal{L}}(q)$, Markov's inequality implies

$$\mathbb{P}\left(h_Q^2(q, p) > \frac{\widehat{\mathcal{L}}(p) - \widehat{\mathcal{L}}(q)}{2} + \frac{\log(K/\delta)}{M}\right) \le \frac{\delta}{K}. \tag{28}$$

A union bound over Π proves equation 22 and therefore equation 23. Finally, Jensen's inequality, the relation between total variation and Hellinger distance, and the marginal density-ratio bound imply

$$\begin{aligned} \mathbb{E}_{R_X} \mathrm{TV}(q, \widehat{p}) &\le \sqrt{\mathbb{E}_{R_X} \mathrm{TV}^2(q, \widehat{p})} \le \sqrt{2 \mathbb{E}_{R_X} h^2(q, \widehat{p} \mid X)} \\ &\le \sqrt{2C h_Q^2(q, \widehat{p})}. \end{aligned} \tag{29}$$

Substitution proves equation 24. □

Lemma G.5 (Approximation error under bounded log loss)**.** *Under the same block sampling scheme and finite class* Π*, suppose there exist constants* $a \in \mathbb{R}$ *and* $B > 0$ *such that, for every* $p \in \Pi$*,*

$$a \le -\log p(A^\star \mid X) \le a + B \qquad Q\text{-almost surely}.$$

Assume also $\mathbb{E}_Q|\log q(A^\star \mid X)| < \infty$*. Let*

$$\epsilon_{\mathrm{app}} := \min_{p\in\Pi} \mathbb{E}_{X\sim Q_X} \mathrm{KL}(q(\cdot \mid X) \,\|\, p(\cdot \mid X)),$$

and suppose $\widehat{p}$ *satisfies* $\widehat{\mathcal{L}}(\widehat{p}) \le \min_{p\in\Pi} \widehat{\mathcal{L}}(p) + \zeta$*. Then, with probability at least* $1-\delta$*,*

$$g_M := \mathbb{E}_{X\sim Q_X} \mathrm{KL}(q(\cdot \mid X) \,\|\, \widehat{p}(\cdot \mid X)) \le \epsilon_{\mathrm{app}} + \zeta + 2B\sqrt{\frac{\log(2K/\delta)}{2M}}. \tag{30}$$

If $R_X \ll Q_X$ *with* $\mathrm{d}R_X/\mathrm{d}Q_X \le C$*, then, on the same event,*

$$\mathbb{E}_{R_X} \mathrm{TV}(q, \widehat{p}) \le \sqrt{\frac{C}{2}\left(\epsilon_{\mathrm{app}} + \zeta + 2B\sqrt{\frac{\log(2K/\delta)}{2M}}\right)}. \tag{31}$$

Here ζ *is an empirical optimization error relative to the best member of* Π*, in contrast to the comparison with* q *in Lemma G.4.*

Proof. Write $\mathcal{L}_Q(p) = \mathbb{E}_Q[-\log p(A^\star \mid X)]$. Each block average in equation 19 belongs to the same interval $[a, a+B]$, and its expectation is $\mathcal{L}_Q(p)$. Hoeffding's inequality and a union bound over the K candidates give, with probability at least $1-\delta$,

$$\sup_{p\in\Pi} |\widehat{\mathcal{L}}(p) - \mathcal{L}_Q(p)| \le u_M, \qquad u_M = B\sqrt{\frac{\log(2K/\delta)}{2M}}.$$

On this event, for a population minimizer $p_{\mathrm{app}} \in \Pi$,

$$\begin{aligned}\mathcal{L}_Q(\widehat{p}) &\le \widehat{\mathcal{L}}(\widehat{p}) + u_M \le \widehat{\mathcal{L}}(p_{\mathrm{app}}) + \zeta + u_M \\ &\le \mathcal{L}_Q(p_{\mathrm{app}}) + \zeta + 2u_M.\end{aligned} \tag{32}$$

The integrability assumptions ensure that all population cross-entropies in this display and $\mathcal{L}_Q(q)$ are finite. The conditional cross-entropy identity therefore gives

$$\mathcal{L}_Q(p) - \mathcal{L}_Q(q) = \mathbb{E}_{Q_X} \mathrm{KL}(q(\cdot \mid X) \| p(\cdot \mid X)).$$

Subtracting $\mathcal{L}_Q(q)$ proves equation 30. Pinsker's inequality, Jensen's inequality, and marginal density-ratio coverage yield

$$\mathbb{E}_{R_X} \mathrm{TV}(q, \widehat{p}) \le \sqrt{\tfrac{1}{2}\mathbb{E}_{R_X} \mathrm{KL}(q\|\widehat{p})} \le \sqrt{C g_M/2},$$

which proves equation 31. □

Extension through a uniform log-density cover. The finite-class argument can be extended when a suitable finite cover is available, without asserting a universal complexity formula for transformers. Fix $\eta > 0$ before observing the blocks. Suppose a possibly infinite policy class Π has a data-independent finite set of conditional densities Π_η of size K_η, such that for every $p \in \Pi$ there is $p_\eta \in \Pi_\eta$ satisfying

$$|\log p(a \mid x) - \log p_\eta(a \mid x)| \le \eta. \tag{33}$$

Require this inequality to hold on a fixed common set of full Q measure, simultaneously for all covered policies, with the log densities finite there. This uniform requirement permits a data-dependent learned policy to use the same cover. If $\widehat{\mathcal{L}}(\widehat{p}) \le \widehat{\mathcal{L}}(q) + \zeta$, then Lemma G.4, applied to Π_η, gives with probability at least $1-\delta$,

$$h_Q^2(q, \widehat{p}) \le \frac{\zeta}{2} + \eta + \frac{\log(K_\eta/\delta)}{M}. \tag{34}$$

To see this, choose the cover element p_η associated with $\widehat{p}$. The cover condition in equation 33 gives $\widehat{\mathcal{L}}(p_\eta) \le \widehat{\mathcal{L}}(q) + \zeta + \eta$, and hence

$$h_Q^2(q, p_\eta) \le (\zeta + \eta)/2 + \log(K_\eta/\delta)/M.$$

The same log-density inequality bounds the affinity from below:

$$1 - h_Q^2(q, \widehat{p}) \ge e^{-\eta/2}[1 - h_Q^2(q, p_\eta)].$$

It follows that $h_Q^2(q, \widehat{p}) \le h_Q^2(q, p_\eta) + 1 - e^{-\eta/2} \le h_Q^2(q, p_\eta) + \eta/2$, proving the claim. Under the marginal coverage assumption, the corresponding total-variation bound is

$$\mathbb{E}_{R_X} \operatorname{TV}(q, \widehat{p}) \le \sqrt{C\zeta + 2C\eta + \frac{2C \log(K_\eta/\delta)}{M}}.$$

A usable rate therefore requires an explicit log-density cover for the specified policy class and input domain; it does not follow merely from using a transformer architecture.

Scope of the sampling and coverage assumptions. The block law may be defined conditional on fixed pretrained experts, curriculum generators, and other preprocessing components. Conditional application of the lemmas requires that, after fixing these components, the blocks remain independent and identically distributed and the hypothesis class remains data-independent. Within-block adaptive adversarial search is permitted. A shared adaptive process that couples different task blocks is not covered by this argument. Similarly, conditioning on components learned jointly from the same blocks does not by itself restore block independence. The effective sample count M is the number of independent blocks, not the number of repeated trajectories from a fixed collection of nominal tasks. Finally, marginal query coverage in these lemmas transfers the learned conditional distribution q to deployment. Transferring an error that also depends on the hidden environment requires coverage of the corresponding *joint* environment–context–query distribution.

G.5 From Average Error to Worst-Case Gain

Lemma G.6 (Uniform control from local curriculum mass)**.** *Let $f : \Psi \to [0, H]$ be L-Lipschitz on a compact metric space. Suppose $P_\Psi(B_r(\psi)) \ge cr^d$ for all ψ and $0 < r \le r_0$, and $\mathbb{E} f \le u$. Then $\sup_\psi f(\psi) \le \mathcal{B}_d(u)$ from equation 13. For $L > 0$ and $d > 0$, the sharper bound*

$$\sup_\psi f(\psi) \le \left(\frac{(d+1)L^d u}{c} \right)^{1/(d+1)} \tag{35}$$

holds whenever $u \le cLr_0^{d+1}/(d+1)$. Without stronger assumptions the exponent $1/(d+1)$ cannot be improved.

Proof. For $y \in B_r(\psi)$, Lipschitz continuity gives $f(\psi) \le f(y) + Lr$. Integrating over the ball, using $f \ge 0$ and dividing by its mass, yields $f(\psi) \le u/(cr^d) + Lr$. Optimize r and use $f \le H$ to obtain the first claim.

For the sharper claim, let $S = \max f$ and let ψ_0 attain the maximum. The cone inequality $f(y) \ge (S - L \operatorname{dist}(y, \psi_0))_+$ and the layer-cake formula give

$$u \ge \int_0^S P_\Psi(f > t)\, dt \ge c \int_0^S \min\{r_0, (S-t)/L\}^d\, dt =: G(S).$$

Strict versus non-strict ball boundaries can be handled by radii increasing to the displayed radius. Direct integration gives

$$G(S) = \begin{cases} cS^{d+1}/((d+1)L^d), & S \le Lr_0, \\ cr_0^d\big(S - \frac{d}{d+1}Lr_0\big), & S > Lr_0. \end{cases}$$

Consequently $u \le cLr_0^{d+1}/(d+1)$ implies $S \le Lr_0$, including the equality case, and inversion proves equation 35. For arbitrary u, $\min\{H, G^{-1}(u)\}$ is another valid uniform bound. If $L = 0$, f is constant and $S \le u$.

To see sharpness, use uniform probability on $[0,1]^d$, Euclidean distance, and $r_0 \le 1/2$. With v_d the volume of the unit Euclidean ball, every such ball intersected with the cube has mass at least cr^d for $c = v_d/2^d$. For $S/L \le r_0$, the corner cone $f(x) = (S - L\|x\|)_+$ has integral exactly $cS^{d+1}/((d+1)L^d)$ and maximum S. It attains both the exponent and the constant in equation 35. □

Interpreting the disturbance dimension. The exponent d describes local curriculum mass on the metric disturbance space; it need not equal the number of weights used to implement an adversary. If disturbances admit coordinates $z \in \mathcal{Z} \subset \mathbb{R}^m$ such that the context-averaged regret depends only on z, is Lipschitz in z, and the induced curriculum satisfies the uniform ball-mass condition with exponent m, Lemma G.6 applies with $d = m$. A mass–friction family can therefore admit $d = 2$ when these coverage and regularity conditions hold. This interpretation concerns the covered disturbance family; a low-dimensional learned latent alone does not establish these conditions.

Proof of Theorems 5.1 and G.1. Lemma G.4 gives $\eta_{\widehat{\theta}} \le \sqrt{C\zeta_M + 2C\log(K/\delta)/M}$ with probability at least $1-\delta$. Substitute it into Proposition G.3 to establish equation 12. The nonnegative regret $f_{\widehat{\theta}}$ lies in $[0, H]$; apply Lemma G.6 on the same event. Finally,

$$\inf_{\psi} \bar{J}_{\widehat{\theta}}(\psi) = \inf_{\psi} \big[J^*(\psi) - f_{\widehat{\theta}}(\psi)\big] \ge \inf_{\psi} J^*(\psi) - \sup_{\psi} f_{\widehat{\theta}}(\psi).$$

Subtract V^{rob} to obtain equation 9. Weak minimax duality gives $\Delta_{expert} \ge 0$, because the fully informed optimal policy class contains Π_0 in every environment. For an actual controller $b \in \Pi_0$, its worst-case value $V_b = \inf_\psi J(b,\psi)$ is at most V^{rob}. Thus the bound relative to b has the larger available gap $\Delta_{expert} + V^{\mathrm{rob}} - V_b$. □

What the comparator means. The comparison specifies both the policy class and its deployment information. History-conditioned robust policies can infer disturbances while their parameters remain fixed. Our result compares conventional policies without BRIDGE's additional context with a policy that uses it, accounting separately for collection in Section G.11. A max-min optimum over a class containing the entire BRIDGE collection-and-deployment strategy, with the same information and interaction budget, upper-bounds the worst-case value of every strategy in that class. The gain therefore concerns the additional context available to the policy; max-min optimization itself permits history-dependent robustness.

G.6 Context Information and Finite-Context Rates

The residual action uncertainty $\mathcal{I}_N$ and the information acquired by the collector, $I(e;\xi)$, characterize distinct quantities. Expert-action KL divergences quantify the former, rather than the latter. The following sufficient condition yields an explicit rate when the collected evidence identifies decision-relevant environmental variation.

Proposition G.7 (An informative-context rate)**.** *Let e lie in a compact subset of Euclidean space. Suppose for all e, e', s, h,*

$$\mathrm{KL}(\pi_e^E(\cdot \mid s,h) \| \pi_{e'}^E(\cdot \mid s,h)) \le L_E \|e - e'\|^2.$$

If an estimator measurable in context satisfies $\mathbb{E}_Q\|e - \widehat{e}_N(\xi)\|^2 \le v/N$, then $\mathcal{I}_N \le 2L_E v/N$. In particular, a sufficient collection condition is N conditionally independent episodes with features Y_i such that $\mathbb{E}[Y_i \mid e] = g(e)$, $\mathbb{E}[\|Y_i - g(e)\|^2 \mid e] \le \sigma^2$, and $\|g(e) - g(e')\| \ge \lambda\|e - e'\|$. If g is continuous, projection of the feature mean onto $g(\Psi)$ yields $v = 4\sigma^2/\lambda^2$.

Proof. Conditional on X, draw e' independently from the posterior of e. Convexity of KL in its second argument gives

$$\mathcal{I}_N \le \mathbb{E}_{Q_X}\mathbb{E}_{e,e'|X}\mathrm{KL}(\pi_e^E \| \pi_{e'}^E) \le L_E \mathbb{E}\|e - e'\|^2 = 2L_E \mathbb{E}\|e - \mathbb{E}[e \mid X]\|^2 \le 2L_E v/N.$$

The last inequality is the optimality of conditional expectation among X-measurable squared-error estimators; $\widehat{e}_N(\xi)$ is one such estimator.

For the sufficient condition, let $\bar{Y} = N^{-1}\sum_i Y_i$ and choose a measurable minimizer $\widehat{e}_N \in \arg\min_{e'} \|g(e') - \bar{Y}\|$. Compactness and continuity provide a minimizer. Projection and the inverse separation imply

$$\lambda\|e - \widehat{e}_N\| \le \|g(e) - g(\widehat{e}_N)\| \le 2\|g(e) - \bar{Y}\|.$$

Conditional independence gives $\mathbb{E}\|g(e) - \bar{Y}\|^2 \le \sigma^2/N$. Squaring the preceding inequality proves the claim. □

The estimator is used only in the analysis; deployment requires neither its computation nor parameter updates. When training and deployment collection laws coincide, the condition applies directly to the robust collector. Otherwise, it concerns the training law, with joint domination controlling deployment mismatch. Common-covariance Gaussian experts with Lipschitz means satisfy the expert-KL condition, whereas differing deterministic continuous actions generally do not. The rate counts N independent collection episodes; applying it to N correlated transitions requires a mixing or martingale argument with an appropriate effective sample size.

Substitution into equation 12 gives the explicit bound

$$\varepsilon_{N,M} \le \epsilon_E + H^2 \left[\sqrt{CL_E v/N} + \sqrt{C\zeta_M + 2C\log(K/\delta)/M}\right].$$

For exact experts, zero optimization slack, and constants uniformly bounded in N, M, Lemma G.6 recovers the claimed $\widetilde{O}((N^{-1/2}+M^{-1/2})^{1/(d+1)})$ error. Horizon, class complexity, context stability, and coverage constants are not universally constant and must be retained when they vary with scale.

Why robust collection alone is insufficient. Consider a one-state, one-step MDP, with ψ uniform on $[-1,-1/2]\cup[1/2,1]$ and actions a_0, a_L, a_R. Rewards are $r(a_0) = 1/2$, $r(a_L) = 1/2 + \psi/2$, and $r(a_R) = 1/2 - \psi/2$. The robust collector always takes a_0, achieving the max-min value $1/2$. Its context is identical for every ψ, no matter how many episodes are collected, and state coverage is complete. A fully informed expert chooses a_L or a_R according to the sign of ψ, with worst-case value $3/4$. Nevertheless, a posterior expert predictor assigns probability $1/2$ to each and attains only $1/2$. Thus $\Delta_{expert} = 1/4$ but improvement is zero even with unlimited pretraining and context. Here $\mathcal{I}_N = \log 2$ for every N. This example separates preservation of the robust baseline during collection from the information needed for subsequent improvement.

G.7 A Value-Based Bound for Continuous Control

The conditional-information bound is useful when uncertainty about expert actions is finite. For deterministic experts in continuous action spaces, however, this information quantity can be infinite even when a good response can be learned. We therefore give an alternative that measures uncertainty through the value of the actions. It requires neither a density for the expert policy nor continuity of the optimal action as the environment changes.

Use the training law Q, deployment law R_θ, and posterior expert predictor q_N from Section G.2. Let $V^*_{e,h}$ and $Q^*_{e,h}$ be the optimal value and action-value functions, respectively.

Proposition G.8 (Value-based adaptation and distillation)**.** *Assume the full joint coverage condition $R_\theta \le CQ_{e,X}$ for some $C \ge 1$. Suppose that*

$$\sup_{s,a,h} |Q^*_{e,h}(s,a) - Q^*_{e',h}(s,a)| \le L_Q d(e,e'), \qquad |Q^*_{e,h}(s,a) - Q^*_{e,h}(s,a')| \le L_a\|a - a'\| \quad (36)$$

for all relevant environments, states, times, and actions. Let

$$\epsilon_E^{\mathrm{adv}} := \mathbb{E}_Q \mathbb{E}_{a\sim\pi^E_{e,h}(\cdot|s)}\left[V^*_{e,h}(s) - Q^*_{e,h}(s,a)\right] \quad (37)$$

be the expert's average optimal advantage on the pretraining query distribution. Assume there exists a context estimator $\widehat{e}_N(\xi)$ satisfying

$$\mathbb{E}_Q d(e, \widehat{e}_N(\xi)) \le \rho_N. \quad (38)$$

Then, whenever the Wasserstein distances below are finite,

$$\boxed{\mathbb{E}_e f_\theta(e) \le CH\big(\epsilon_E^{\mathrm{adv}} + 4L_Q\rho_N\big) + HL_a\,\mathbb{E}_{R_\theta} W_1(q_N(\cdot \mid X), T_\theta(\cdot \mid X))\,.} \quad (39)$$

The estimator in equation 38 is an analysis device: the deployed model need not explicitly compute it or update any parameters.

Proof. Define the nonnegative optimal advantage $D_{e,h}(s,a) := V^*_{e,h}(s) - Q^*_{e,h}(s,a)$. Taking the supremum over actions in equation 36 shows that $V^*_{e,h}(s)$ is also L_Q-Lipschitz in e. Consequently,

$$D_{e,h}(s,a) \le D_{e',h}(s,a) + 2L_Q d(e,e'). \tag{40}$$

Under Q, draw e' independently from the posterior $Q(\cdot \mid X)$. Conditional on X, the true e and this posterior draw have the same distribution. The triangle inequality therefore gives

$$\begin{aligned}\mathbb{E}d(e,e') &\le \mathbb{E}d(e,\widehat{e}_N(\xi)) + \mathbb{E}d(e',\widehat{e}_N(\xi)) \\ &= 2\mathbb{E}_Q d(e,\widehat{e}_N(\xi)) \le 2\rho_N.\end{aligned} \tag{41}$$

Average equation 40 over this posterior draw and then over $a \sim \pi^E_{e',h}(\cdot \mid s)$. Posterior symmetry identifies the first term on the right with equation 37, yielding

$$\mathbb{E}_Q \mathbb{E}_{a\sim q_N(\cdot|X)} D_{e,h}(s,a) \le \epsilon^{\mathrm{adv}}_E + 4L_Q\rho_N. \tag{42}$$

This step uses continuity of values; it does not require nearby environments to have the same optimal action.

For any coupling of $a \sim q_N(\cdot \mid X)$ and $a' \sim T_\theta(\cdot \mid X)$, action-value Lipschitzness implies

$$\mathbb{E}D_{e,h}(s,a') \le \mathbb{E}D_{e,h}(s,a) + L_a\mathbb{E}\|a - a'\|. \tag{43}$$

Taking the infimum over couplings gives the same inequality with the last expectation replaced by $W_1(q_N, T_\theta)$. The finite-horizon performance-difference identity and the definition of R_θ imply

$$\mathbb{E}_e f_\theta(e) = H\mathbb{E}_{R_\theta}\mathbb{E}_{a\sim T_\theta(\cdot|X)} D_{e,h}(s,a). \tag{44}$$

The posterior advantage is nonnegative, so joint coverage allows its expectation under R_θ to be bounded by C times its expectation under Q. Substituting equation 42 proves equation 39. □

Obtaining the environment-value regularity. The first condition in equation 36 follows from standard regularity of the underlying MDP. Suppose rewards and transition kernels satisfy

$$|r_e(s,a) - r_{e'}(s,a)| \le L_r d(e,e'), \qquad \mathrm{TV}(P_e(\cdot \mid s,a), P_{e'}(\cdot \mid s,a)) \le L_P d(e,e'). \tag{45}$$

Let B_h be a Lipschitz constant for $Q^*_{e,h}$ in e, uniformly over states and actions. The Bellman recursion, $0 \le V^*_{e,h+1} \le H-h$, and the total-variation bound imply $B_h \le L_r + B_{h+1} + (H-h)L_P$, with $B_{H+1} = 0$. Backward induction yields

$$B_h \le (H-h+1)L_r + \frac{(H-h)(H-h+1)}{2} L_P. \tag{46}$$

Thus $L_Q = HL_r + H(H-1)L_P/2$ is valid. Action-value Lipschitzness is a separate condition; it is not implied by regularity only in the environment parameter.

The collection condition of Proposition G.7 also gives $\rho_N \le 2\sigma/(\lambda\sqrt{N})$ by Jensen's inequality, without requiring its expert-KL smoothness assumption. Thus this value-based route retains a finite-context rate when optimal actions change discontinuously.

Connection to supervised continuous-action losses. Applying Proposition G.8 to a continuous-action implementation requires its training loss to control a value-relevant discrepancy between the learned and posterior expert-action distributions. For a deterministic predictor $T_\theta(\cdot \mid X) = \delta_{u_\theta(X)}$, define the population squared action error

$$\mathcal{L}_{\mathrm{MSE}}(\theta) = \mathbb{E}_Q \mathbb{E}_{A^*\sim\pi^E_{e,h}(\cdot|s)} \|A^* - u_\theta(X)\|_2^2. \tag{47}$$

Because $W_1(q_N, \delta_u) = \mathbb{E}_{A\sim q_N}\|A - u\|_2$, Jensen's inequality and the joint coverage condition give

$$\mathbb{E}_{R_\theta} W_1(q_N, \delta_{u_\theta}) \le \sqrt{C\mathcal{L}_{\mathrm{MSE}}(\theta)}. \tag{48}$$

This yields a valid substitute for the final term in equation 39. However, it is the *total* population squared error that appears. If $\bar{a}(X) = \mathbb{E}[A^* \mid X]$, the regression decomposition is

$$\mathcal{L}_{\mathrm{MSE}}(\theta) = \mathbb{E}\|A^* - \bar{a}(X)\|_2^2 + \mathbb{E}\|\bar{a}(X) - u_\theta(X)\|_2^2. \tag{49}$$

At zero excess regression error, the second term vanishes while the first retains uncertainty in the expert-action distribution, including multimodality. Predicting the conditional mean therefore need not reproduce that distribution. A finite-sample guarantee for a continuous-action implementation must control generalization under its actual loss and retain the corresponding residual or distributional approximation error.

An optional likelihood-based version. The range of the optimal advantage is at most H, so total variation can replace the action-Lipschitz argument. Under Lemma G.4, on its probability-$1-\delta$ event,

$$\mathbb{E}_e f_{\widehat{\theta}}(e) \le CH\big(\epsilon_E^{\mathrm{adv}} + 4L_Q\rho_N\big) + H^2\sqrt{C\zeta_M + 2C\log(K/\delta)/M}. \tag{50}$$

This avoids action mutual information, but still requires the conditional-likelihood assumptions. A deterministic continuous-action regressor does not automatically satisfy them. Either average-regret bound can be substituted into the worst-case and unseen-task results when their additional coverage and context-aware regularity conditions hold.

G.8 Regularity of Context-Conditioned Returns

Smooth MDP parameters alone do not imply an N-uniform regularity bound on a policy whose context distribution changes with the environment. The following sufficient condition accounts for both effects.

Lemma G.9 (Context-aware regret regularity)**.** *Suppose $J^*(e)$ is L_*-Lipschitz in e, and every fixed-context policy has an L_J-Lipschitz return as e varies. Let ρ be a metric on contexts. Assume*

$$W_1(B_e^N, B_{e'}^N) \le L_B \operatorname{dist}(e, e'), \qquad |J(T_\theta(\xi), e) - J(T_\theta(\xi'), e)| \le L_T\rho(\xi, \xi').$$

Then f_θ is Lipschitz with constant $L = L_* + L_J + L_T L_B$.

Proof. Add and subtract $\mathbb{E}_{B_e^N} J(T_\theta(\xi), e')$ when comparing $\bar{J}_\theta(e)$ and $\bar{J}_\theta(e')$. The fixed-context return bound controls one difference by $L_J \operatorname{dist}(e, e')$. Coupling the two context laws controls the other by $L_T W_1(B_e^N, B_{e'}^N)$. Add the L_* bound for the optimal return. □

For independent context episodes, an averaged product metric $\rho(\xi, \xi') = N^{-1}\sum_i \rho_1(\xi_i, \xi_i')$ and a Lipschitz one-episode context law yield an N-independent L_B by product coupling. An N-uniform L_T remains a policy-stability assumption. If it is unavailable, write L_N explicitly: the sufficient condition for the error to vanish is $L_N^d \varepsilon_{N,M} \to 0$, with the small-radius condition also satisfied. A lower density on an arbitrary compact support is likewise insufficient for the uniform ball-mass assumption; the support's local geometry matters.

G.9 Extension to Unseen Nominal Tasks

Let nominal tasks follow a population μ, with disturbance curriculum ν_τ for task τ. Pretraining observes independent nominal-task collections, and deployment draws a fresh task independently from μ_{test}. Apply the average regret bound equation 12 to the joint population $\mu(d\tau)\nu_\tau(d\psi)$, using the environment index $e = (\tau, \psi)$. Define $V^{\mathrm{rob}}(\tau)$ and $\Delta_{expert}(\tau)$ taskwise as in equation 8.

Corollary G.10 (Robust generalization to unseen tasks)**.** *Apply the average regret bound to $\mu(d\tau)\nu_\tau(d\psi)$. If the theorem's neighborhood-mass and regret-regularity conditions hold uniformly within tasks, with disturbance dimension d_ψ, and the fresh-task distribution satisfies $\mu_{\mathrm{test}} \le \kappa\mu$, then, on the same event,*

$$\mathbb{E}_{\tau\sim\mu_{\mathrm{test}}}\left[\inf_\psi \bar{J}_{\widehat{\theta}}(\tau, \psi) - V^{\mathrm{rob}}(\tau)\right] \ge \mathbb{E}_{\tau\sim\mu_{\mathrm{test}}}\Delta_{expert}(\tau) - \mathcal{B}_{d_\psi}(\kappa\varepsilon_{N,M}). \tag{51}$$

For $e = (\tau, \psi)$, the conditional mutual-information chain rule gives

$$\mathcal{I}_N = I_Q(A^\star; \tau \mid X) + I_Q(A^\star; \psi \mid X, \tau). \tag{52}$$

The first term measures task uncertainty relevant to choosing an action; the second measures disturbance uncertainty that remains relevant after the task is known. Both contribute to the context term in equation 12.

Proof of Corollary G.10. Let $a(\tau) = \mathbb{E}_{\psi\sim\nu_\tau} f_{\widehat{\theta}}(\tau,\psi)$. The joint average theorem gives $\mathbb{E}_\mu a \le \varepsilon_{N,M}$. Since $a \ge 0$ and $\mu_{\text{test}} \le \kappa\mu$, we have $\mathbb{E}_{\mu_{\text{test}}} a \le \kappa\varepsilon_{N,M}$. Within each task, Lemma G.6 yields

$$\inf_\psi \bar{J}_{\widehat{\theta}}(\tau,\psi) - V^{\text{rob}}(\tau) \ge \Delta_{expert}(\tau) - \mathcal{B}_{d_\psi}(a(\tau)).$$

The function $\mathcal{B}_{d_\psi}$ is nondecreasing and concave: it is the infimum of affine nondecreasing functions of its argument, together with the constant H. Jensen's inequality then gives $\mathbb{E}_{\mu_{\text{test}}}\mathcal{B}_{d_\psi}(a) \le \mathcal{B}_{d_\psi}(\kappa\varepsilon_{N,M})$, proving the result. The information decomposition equation 52 is the conditional mutual-information chain rule. □

For $\mu_{\text{test}} = \mu$, $\kappa = 1$, and the guarantee concerns an independent nominal-task draw from the population. With a nonatomic μ, this draw is absent from any finite nominal training set almost surely; in a discrete task space, it may repeat a training task. Applying the corollary to a prescribed disjoint train/test split requires corresponding sampling or transfer assumptions. Correlated trajectories from a fixed task pool leave unchanged the number of independent nominal-task draws used in the learning lemma.

Corollary G.11 (Uniform extension over the task space)**.** *Suppose the joint task–disturbance space has uniform ball mass cr^D and $f_{\widehat{\theta}}(\tau,\psi)$ is jointly L-Lipschitz. Under the average theorem's assumptions, with the same probability, every nominal task in this covered space satisfies*

$$\inf_\psi \bar{J}_{\widehat{\theta}}(\tau,\psi) - V^{\text{rob}}(\tau) \ge \Delta_{expert}(\tau) - \mathcal{B}_D(\varepsilon_{N,M}).$$

Proof. Apply Lemma G.6 on the joint space and then subtract the per-task robust value. Unlike Corollary G.10, this statement controls the supremum over nominal tasks as well, and its dimension is correspondingly that of the joint space. □

If the undisturbed environment $\psi = 0$ belongs to each covered uncertainty set, the same uniform regret bound controls it. This provides an absolute-performance guarantee there; it does not establish that adversarial augmentation necessarily improves performance over every nominal-only pretraining procedure. The observed undisturbed generalization benefit is an empirical comparison in Section 6. No claim of arbitrary extrapolation beyond the covered task population follows from these corollaries.

G.10 ADAPT and Retention of Decision-Relevant Information

For a fixed representation $Z = g(X)$ containing the two latents supplied to ADAPT's decoder, the prediction error decomposes as

$$\begin{aligned}&\mathbb{E}_Q \mathrm{KL}(q_N(\cdot \mid X) \| T_\theta(\cdot \mid Z)) \\ &\quad = I_Q(A^\star; X \mid Z) + \mathbb{E}_Q \mathrm{KL}(Q(A^\star \mid Z) \| T_\theta(\cdot \mid Z)).\end{aligned} \tag{53}$$

The first term is decision-relevant information lost by the representation, and the second is error in decoding the retained information. Their sum controls the prediction contribution to return error through Proposition G.3.

Proposition G.12 (Representation and decoding error)**.** *For any fixed measurable representation $Z = g(X)$ and decoder $p_\theta(\cdot \mid Z)$, the decomposition equation 53 holds, whenever the divergences are finite. For an encoder learned from data, it holds conditionally on each realized encoder, with expectations over an independent population example.*

Proof. Set $q_Z = Q(A^\star \mid Z)$, and add and subtract $\log q_Z(A^\star \mid Z)$:

$$\begin{aligned}\mathbb{E}\log\frac{q_N(A^\star \mid X)}{p_\theta(A^\star \mid Z)} &= \mathbb{E}\log\frac{q_N(A^\star \mid X)}{q_Z(A^\star \mid Z)} + \mathbb{E}\log\frac{q_Z(A^\star \mid Z)}{p_\theta(A^\star \mid Z)} \\ &= I(A^\star; X \mid Z) + \mathbb{E}_Z \mathrm{KL}(q_Z \| p_\theta).\end{aligned}$$

The first equality in the second line uses the fact that Z is a function of X. Conditioning on training data fixes a learned encoder and leaves the same identity for a fresh example. The identity alone is not a generalization bound for a data-dependent representation; such a bound must control the full encoder–decoder class. □

For ADAPT, Z contains the task and adversary latents; the query state enters through the encoder, as in Algorithm 2. Task anchoring and conditioned dynamics prediction encourage retention of different useful evidence; the action objective trains their joint use. The information loss $I(A^\star; X \mid Z)$ formalizes the cost of discarding decision-relevant evidence. It does not require the latents to identify the entire task or disturbance. Conversely, the raw-context quantity $I(A^\star; e \mid X)$ is a property of the expert/data law and cannot be decreased by changing the encoder alone.

Orthogonality is geometric, not statistical independence. For example, let $u(\tau)$ and $v(\psi)$ encode the two factors and take $z_\tau = (u(\tau), v(\psi), 0, 0)$ and $z_\phi = (0, 0, u(\tau), v(\psi))$. These latents have zero inner product while both encode both factors and support exact task and dynamics prediction. Stop-gradient alters the routing of the training objective and leaves this representational possibility intact. The auxiliary objectives therefore provide inductive biases whose effects are evaluated through return ablations and representation probes; minimizing them alone does not establish deterministic task identification or causal disentanglement.

G.11 Robust Context Collection and the Deployment Horizon

Proposition G.13 (Preserving the baseline during collection)**.** *In a fixed environment, suppose K_0 reset episodes collect context by executing the actual robust baseline b, followed by L_0 reset episodes using a frozen predictor with that same collected context held fixed. During collection, every measurable trajectory event has exactly the same probability as under b. Let $\bar{J}_\theta(e)$ be the expected return per subsequent episode under the resulting context. The average return over the full horizon is*

$$F(e) = \frac{K_0 J(b, e) + L_0 \bar{J}_\theta(e)}{K_0 + L_0}.$$

Writing $V_b = \inf_e J(b, e)$, its worst-case gain satisfies

$$\inf_e F(e) - V_b \geq \frac{L_0}{K_0 + L_0} \left[\inf_e \bar{J}_\theta(e) - V_b \right].$$

Proof. The collector uses the same action law, transitions, and initial distribution as b, so its trajectory law is identical. Conditional on the collected context, each subsequent episode has the corresponding context-conditioned return. Averaging over that context and adding the expected collection return gives the displayed expression for F. Finally, the infimum of a sum is at least the sum of its infima. Subtract V_b to obtain the gain inequality. □

The collection guarantee preserves the risk of the same baseline over the same collection period, including any risk inherent to that baseline. Post-collection improvement requires the derived error to be smaller than the relevant oracle gap. A within-episode handoff requires continuation values from the reached state and history, rather than reset-episode returns. For unseen nominal tasks, the proposition requires a baseline actually available for context collection; the robust comparator in Corollary G.10 alone does not provide such a controller.