SUAF: SPECTRAL UNIVERSAL ADVANTAGE FEATURES FOR ZERO-SHOT POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward-independent representations enable reinforcement learning agents to reuse knowledge across tasks, but limited representation capacity raises three fundamental questions: what statistical dependence should these representations capture, which components of that dependence should be retained for policy improvement, and how does discarding the remaining components affect decision-making? We address these questions through canonical dependence analysis. For a reference policy, we construct an action-centered, positive-lag canonical dependence matrix (CDM) that isolates the additional predictive contribution of the current action beyond the current state. Its singular decomposition yields Spectral Universal Advantage Features (SUAF), paired features that recover delayed advantage through linear reward projections. We show that the spectrum quantifies action-incremental χ^2 dependence and determines reward-specific approximation errors. Retaining the leading modes gives optimal rank-constrained approximations under both isotropic average-case and reward-uniform minimax criteria. Separating the immediate reward contrast also avoids the spectral floor arising from direct compression of the full advantage operator. To learn SUAF from reward-free data, we develop an action-centered Soft-HGR objective with bidirectional temporal-difference estimation. We further connect spectral approximation errors to sufficient conditions for policy improvement. Finite-state experiments validate the spectral predictions and feature-learning method, while zero-shot evaluations on discrete tasks and continuous-control benchmarks demonstrate the utility of the learned representations across rewards.

1 INTRODUCTION

Recent advances in large language models (LLMs) and embodied artificial intelligence (AI) have triggered a growing ambition to build general-purpose agents that operate in open-ended environments (Xi et al., 2025; Liu et al., 2025). Unlike traditional task-specific models, these agents are expected to handle a diverse array of goals, costs, and preferences within the same environment, often specified on the fly by human instructions or changing task contexts. Although these changes alter the reward function, knowledge of how actions influence future outcomes remains reusable. Reward-independent representations seek to capture this knowledge and support new tasks without learning a separate critic from scratch.

Among reward-independent representations, successor features (SFs) provide a foundational framework for such reuse by separating reward weights from the discounted accumulation of features (Dayan, 1993; Barreto et al., 2017b). Building on this, more recent successor-measure (SM) representations seek to encode behavior across policies using reward-free interaction data (Blier et al., 2021; Touati & Ollivier, 2021; Touati et al., 2023). However, a critical challenge arises: which features should be learned? To address this problem, Ollivier (2025) connects optimal features to the spectral structure of the full advantage operator through expected downstream performance under strong behavior-cloning regularization. Despite these theoretical advances, limited representation capacity still raises three fundamental questions: what statistical dependence should these representations capture, which components of that dependence should be retained for policy improvement, and how does discarding the remaining components affect decision-making?

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.

To address these problems, we introduce **Spectral Universal Advantage Features** (SUAF), a reward-independent representation of delayed advantage with respect to the corresponding behavior policy. Our starting point is that future outcomes depend on both state and action, while advantage measures the action's contribution relative to a reference policy. We therefore isolate the dependence that remains after subtracting the policy-averaged future visitation. Building on canonical dependence analysis (Huang et al., 2024), we construct an action-centered, positive-lag canonical dependence matrix (CDM). Its singular decomposition yields paired features of current state-action pairs and future outcomes, whose full expansion recovers delayed advantage through linear reward projections. A downstream reward enters linearly through the future features, recovering delayed advantage. To determine which components to retain, we separate the immediate reward contrast from delayed effects. The immediate term induces a nonzero spectral floor, limiting direct low-rank compression and when evaluable, SUAF keeps it explicit and allocates capacity to the delayed component. We further find that the squared singular values quantify action-incremental χ^2- dependence, and retaining leading modes gives an optimal rank-constrained approximation of delayed advantage under both average and worst-case reward objectives. We then ask how discarding the remaining components affects decision-making. The error depends on how a reward aligns with retained features. We derive an exact approximation-error expression via projections onto discarded modes, distinguishing aggregate lost dependence from task-specific information loss, and obtain a sufficient condition under which the predicted policy-improvement gain exceeds the approximation penalty. We further bound local representation reuse under policy changes within trust region. To learn SUAF without constructing the CDM, we develop an action-centered Soft-HGR objective with bidirectional temporal-difference estimation, using forward and backward predictors from reward-free transitions. The learned representation supports advantage readout and policy extraction. Experiments examine the spectral floor, truncation errors, and feature-learning accuracy, with zero-shot evaluations on discrete and continuous-control benchmarks.

In summary, our contributions are threefold:

1. We introduce SUAF through an action-centered positive-lag CDM, connecting delayed advantage to the incremental dependence between current actions and future outcomes. We identify a spectral floor in the full advantage operator that motivates retaining the immediate reward contrast explicitly and compressing only the delayed component.
2. We establish average-case and reward-uniform minimax optimality for spectral approximation of delayed advantage. We relate the retained spectrum to action-incremental χ^2 dependence and derive reward-specific truncation errors that yield sufficient conditions for policy improvement.
3. We develop an action-centered Soft-HGR learning method with bidirectional TD estimation. Experiments validate the spectral analysis and feature-learning accuracy, and evaluate zero-shot policy improvement across rewards in discrete and continuous environments.

2 PRELIMINARIES

2.1 UNIVERSAL FEATURE FUNCTIONS FOR HIGH DIMENSIONAL LEARNING

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two random variables with joint distribution P_{XY} and marginal distributions P_X and P_Y. Assuming that $P_X(x) > 0$ and $P_Y(y) > 0$ on their supports, the canonical dependence matrix (CDM) $\widetilde{\boldsymbol{B}} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is defined as

$$\widetilde{\boldsymbol{B}}(y,x) = \frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}.$$

The CDM is the normalized residual between the joint distribution P_{XY} and the independent distribution $P_X P_Y$. In particular, $\widetilde{\boldsymbol{B}} = \mathbf{0}$ if and only if X and Y are independent. When X and Y are weakly dependent, $\|\widetilde{\boldsymbol{B}}\|_{\mathrm{F}}^2 \approx 2I(X;Y)$, where $I(X;Y)$ is mutual information. Let the singular value decomposition (SVD) of the CDM be $\widetilde{\boldsymbol{B}} = \sum_{i=1}^{R} \sigma_i \psi_i^Y \left(\psi_i^X\right)^{\mathrm{T}}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R > 0$, where $R \leq \min\{|\mathcal{X}|-1, |\mathcal{Y}|-1\}$. The corresponding feature functions are defined as

$$f_i(x) = \frac{\psi_i^X(x)}{\sqrt{P_X(x)}}, \qquad g_i(y) = \frac{\psi_i^Y(y)}{\sqrt{P_Y(y)}}.$$

These feature functions are centered and orthonormal under their respective marginal distributions. For any $k \leq R$, the first k pairs of $\{f, g\}$ solve the generalized canonical correlation problem

$$\begin{aligned} \max_{\{f_i, g_i\}_{i=1}^k} \quad & \sum_{i=1}^k \mathbb{E}[f_i(X) g_i(Y)] \\ \text{s.t.} \quad & \mathbb{E}[f_i(X)] = \mathbb{E}[g_i(Y)] = 0, \mathbb{E}[f_i(X) f_j(X)] = \mathbb{E}[g_i(Y) g_j(Y)] = \mathbf{1}(i = j), \end{aligned} \tag{1}$$

whose optimal value is $\sum_{i=1}^k \sigma_i$. The case $k = 1$ reduces to the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation. Under the local information-geometric formulation of (Huang et al., 2024), these leading feature functions are universal in the sense that they optimize the average local inference performance for unknown weak attributes of Y.

2.2 Markov Decision Processes

An MDP is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, T, r, \gamma, \mu \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $T : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ denotes the transition probability with $\mathcal{P}(\cdot)$ denotes a probability distribution over a set, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward function with $|r(s, a)| \leq R_{\max}$ for all (s, a), $\gamma \in [0, 1)$ denotes the discount factor and μ denotes the distribution of initial states. The goal of RL is to learn a policy $\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ with parameters θ that maximize the expected return $J_r(\pi_\theta) = \mathbb{E}_{s_0 \sim \mu} \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)\right]$ where s_t and a_t denotes the state and action at timestep t.

The marginal occupancy measure, $d^\pi(s, a)$ is defined as the normalized probability of visiting state-action pair (s, a) following the policy π from a state sampled from the initial state distribution, which is $d^\pi(s, a) = (1 - \gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}_\pi(s_t = s, a_t = a)$. The successor measure, $M^\pi(s, a, s^+, a^+)$, is defined as the probability of being in state s^+ and taking action a^+ when starting from the state-action pair (s, a) and following the policy π, which is $M^\pi(s^+, a^+|s, a) = (1 - \gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}_\pi(s_t = s^+, a_t = a^+|s_0 = s, a_0 = a)$. The marginal occupancy measure is the expectation of successor measure, namely, $d^\pi(s, a) = \mathbb{E}_{s_0 \sim \mu, a_0 \sim \pi(\cdot|s_0)}[M^\pi(s, a|s_0, a_0)]$. The Q-function defined using successor measure is $Q^\pi(s, a) = \mathbb{E}_{(s^+, a^+) \sim M^\pi}[r(s^+, a^+)|s, a]$ or $Q^\pi = (1 - \gamma)^{-1} M^\pi r$. The V-function is defined as $V^\pi(s) = \sum_a \pi(a|s) Q^\pi(s, a)$ and advantage function is defined as $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$.

3 Spectral Universal Advantage Feature Functions

Definition and Properties of SUAF We establish the SUAF with dependent CDM. To begin with, we define the positive-lag successor measure $M^{+,\pi}(s, a, s^+, a^+)$ as

$$M^{+,\pi}(s^+, a^+|s, a) = (1 - \gamma) \sum_{t=1}^\infty \gamma^{t-1} \mathbb{P}_\pi(s_t = s^+, a_t = a^+|s_0 = s, a_0 = a).$$

This measure represents the normalized discounted distribution of future state–action pairs (s^+, a^+) reached under policy π from (s, a), excluding the current time step. Unlike the marginal occupancy measure d^π, it remains dependent on the initial state–action pair and therefore characterizes how the current action shapes the strictly future visitation distribution. Based on the positive-lag successor measure, we consider the positive-lag dependent CDM (DCDM)[1] as

$$\widetilde{\boldsymbol{B}}^{+,\pi}((s, a), (s^+, a^+)) = \sqrt{d^\pi(s, a)} \frac{M^{+,\pi}(s^+, a^+|s, a) - \mathbb{E}_{a' \sim \pi(\cdot|s)} \left[M^{+,\pi}(s^+, a^+|s, a')\right]}{\sqrt{\nu^{+,\pi}(s^+, a^+)}}, \tag{2}$$

where $\nu^{+,\pi}(\cdot) = \sum_{s,a} d^\pi(s, a) M^{+,\pi}(\cdot \mid s, a)$ is the marginal distribution of (s^+, a^+). The DCDM is the normalized residual between successor measure and the corresponding distribution conditioned on the state alone. In particular, conditioning on s removes the predictive information already contained in the current state, so that $\widetilde{\boldsymbol{B}}^{+,\pi}$ captures the additional influence of the current action a on the strictly future state-action distribution. Especially, when (S^+, A^+) is weakly dependent on (S, A), $\|\widetilde{\boldsymbol{B}}^{+,\pi}\|_{\mathrm{F}}^2 \approx 2I(A; S^+, A^+|S)$ holds[2]. The weak dependence happens when the effective discount horizon $1/(1 - \gamma)$ is substantially longer than the mixing time of the Markov chain

[1] We remove the $t = 0$ term to control the spectrum. See details in Theorem 3.3 and Appendix D.2.

[2] $I(A; S^+, A^+|S)$ denotes the conditional mutual information. See in Appendix A.3 for details.

induced by π. A sufficiently exploratory policy and overlapping stochastic transitions will further weaken the dependence. With the SVD of DCDM as $\widetilde{\boldsymbol{B}}^{+,\pi} = \sum_{i=1}^{R} \sigma_i^{+,\pi} \psi_{i,+,\pi}^{S,A} (\psi_{i,+,\pi}^{S^+,A^+})^{\mathrm{T}}$ with $\sigma_1^{+,\pi} \geq \cdots \geq \sigma_R^{+,\pi} > 0$, where $R \leq |\mathcal{S}| - 1$ indicating a low-rank matrix[3], we define the SUAF as

$$f_i(s,a) = \frac{\psi_{i,+,\pi}^{S,A}(s,a)}{\sqrt{d^\pi(s,a)}}, \quad g_i(s^+,a^+) = \frac{\psi_{i,+,\pi}^{S^+,A^+}(s^+,a^+)}{\sqrt{\nu^{+,\pi}(s^+,a^+)}}. \tag{3}$$

The proposed SUAF shares the properties with the universal features proposed by Huang et al. (2024), and see Appendix A.4 for details.

Reward-Independent Universal Optimality of SUAF In this paragraph, we propose that the SUAF is universal for downstream tasks. Furthermore, when the advantage operator is approximated with rank at most k, we prove that SUAF is min-max optimal for downstream tasks.

Theorem 3.1 (Universal advantage function representation)**.** *For any reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and for every (s,a) in the support of d^π, the advantage function admits the exact representation*

$$A_r^\pi(s,a) = C_\pi r(s,a) + \frac{\gamma}{1-\gamma} \sum_{i=1}^{R} \sigma_i^{+,\pi} f_i(s,a) \langle g_i, r \rangle_{\nu^{+,\pi}}, \tag{4}$$

where $C_\pi r(s,a) := r(s,a) - \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]$ is the immediate centered reward and $\langle g_i, r\rangle_{\nu^{+,\pi}} = \sum_{s,a} \nu^{+,\pi}(s,a) g_i(s,a) r(s,a)$ is the inner product on Hilbert space $L^2(\nu^{+,\pi})$.

Denote the positive-lag future operator as

$$\mathcal{T}^{+,\pi} r(s,a) = \sum_{s^+,a^+} \left(M^{+,\pi}(s^+,a^+ \mid s,a) - \mathbb{E}_{a'\sim\pi(\cdot|s)} \left[M^{+,\pi}(s^+,a^+ \mid s,a') \right] \right) r(s^+,a^+),$$

and the advantage function can be written as $A_r^\pi = C_\pi r + \frac{\gamma}{1-\gamma} \mathcal{T}^{+,\pi} r$ with $\mathcal{T}^{+,\pi} = \sum_i \sigma_i^{+,\pi} f_i \otimes g_i$. Consequently, the DCDM provides a reward-independent representation of the delayed component of the advantage function. The functions $\{g_i\}_{i=1}^R$ form an orthonormal basis for the reward subspace relevant to the delayed advantage. In particular, every reward can be decomposed as $r = \sum_{i=1}^R \langle g_i, r\rangle_{\nu^{+,\pi}} g_i + r_\perp$, where the residual $r_\perp$ satisfies $\mathcal{T}^{+,\pi} r_\perp(s,a) = 0$ for every (s,a). Hence, $r_\perp$ has no effect on the delayed advantage, and the coefficients $\{\langle g_i, r\rangle_{\nu^{+,\pi}}\}_{i=1}^R$ completely characterize the reward information relevant to future credit assignment.

When only $k < R$ features can be used, the approximation is denoted as $\widehat{\mathcal{T}}^{+,\pi}$ with $\mathrm{rank}(\widehat{\mathcal{T}}^{+,\pi}) \leq k$. The corresponding advantage function is $\hat{A}_r^\pi = C_\pi r + \frac{\gamma}{1-\gamma} \widehat{\mathcal{T}}^{+,\pi} r$. Theorem 3.2 then shows that, conditional on the environment and policy, SUAF provides universal, reward-independent advantage features with optimal rank-k approximation risk across downstream tasks.

Theorem 3.2 (Average-case optimality of the rank-k representations)**.** *Let reward function r be sampled from uniform distribution over the unit sphere of $L^2(\nu^{+,\pi})$, and let $m = |\operatorname{supp}(\nu^{+,\pi})|$. Among all reward-independent rank-k approximations $\widehat{\mathcal{T}}^{+,\pi}$ for $\hat{A}_r^\pi$, the top-k truncated representation*

$$A_{r,k}^\pi = C_\pi r + \frac{\gamma}{1-\gamma} \sum_{i=1}^{k} \sigma_i^{+,\pi} f_i \langle g_i, r\rangle_{\nu^{+,\pi}} \tag{5}$$

minimizes the expected advantage approximation error, i.e.,

$$\mathbb{E}_r \left[\left\| A_r^\pi - A_{r,k}^\pi \right\|_{d^\pi}^2 \right] = \inf_{\mathit{rank}(\widehat{\mathcal{T}}^{+,\pi}) \leq k} \mathbb{E}_r \left[\left\| A_r^\pi - \hat{A}_r^\pi \right\|_{d^\pi}^2 \right] = \frac{1}{m} \left(\frac{\gamma}{1-\gamma} \right)^2 \sum_{i>k} \left(\sigma_i^{+,\pi} \right)^2,$$

and top-k truncation is also optimal under minimax case, i.e.,

$$\sup_{\|r\|_{\nu^{+,\pi}} \leq 1} \left\| A_r^\pi - A_{r,k}^\pi \right\|_{d^\pi}^2 = \inf_{\mathit{rank}(\widehat{\mathcal{T}}^{+,\pi}) \leq k} \sup_{\|r\|_{\nu^{+,\pi}} \leq 1} \left\| A_r^\pi - \hat{A}_r^\pi \right\|_{d^\pi}^2 = \left(\frac{\gamma}{1-\gamma} \right)^2 \left(\sigma_{k+1}^{+,\pi} \right)^2.$$

[3]The rank is less than the state-action joint space size. See proof in Appendix A.3.

The uniform distribution over the unit sphere of $L^2(\nu^{+,\pi})$ serves as a task-agnostic reward model, with each reward defining a downstream task.

Furthermore, we claim that the exact immediate term also removes a spectral floor that arises when approximating the full advantage operator.

Theorem 3.3 (Removing the immediate-term spectral floor)**.** *Consider a finite MDP with $0 < \gamma < 1$ and strictly positive d^π and $\nu^{+,\pi}$. Let $\alpha_\pi = \min_s \frac{d^\pi_S(s)}{\nu^{+,\pi}_S(s)}$, $q = |\mathcal{S}|(|\mathcal{A}| - 1)$. For every $0 \le k < q$,*

$$\inf_{\mathrm{rank}(K)\le k} \sup_{\|r\|_{\nu^{+,\pi}}\le 1} \|A^\pi_r - Kr\|^2_{d^\pi} \ge \alpha_\pi + \left(\frac{\gamma}{1-\gamma}\right)^2 \left(\sigma^{+,\pi}_{k+1}\right)^2, \tag{6}$$

where singular values beyond the rank of $\mathcal{T}^{+,\pi}$ are defined to be zero. In contrast, retaining $C_\pi r$ exactly and truncating only the future term incurs worst-case error without α_π as shown in Theorem 3.2. If $\nu^{+,\pi} = d^\pi$, then $\alpha_\pi = 1$ and equation 6 holds with equality.

Therefore, directly compressing the full advantage operator retains a nonzero error floor for $k < q$, whereas the SUAF approximation error follows only the delayed spectrum and vanishes once all nonzero future modes are retained. The improvement comes from computing the known immediate term exactly, rather than spending the learned rank budget on representing it. For per-reward analysis, please refer to Appendix A.5.

4 METHODS

In this section, we use the truncated SUAF to achieve policy improvement. We propose two classes of improvements, including one-step update and trust region update. After that, we introduce methods for estimating SUAF.

One Step Policy Update Following One-step FB (Zheng et al., 2026), we use the SUAF of a fixed behavior policy π to perform reward-conditioned policy improvement. Given a downstream reward r, we compute its coefficients $\beta_i(r) = \langle g^\pi_i, r\rangle_{\nu^{+,\pi}}$ and construct $A^\pi_{r,k}(s,a)$ following equation 5. The improved policy selects actions maximizing this score: $\pi^+_r(s) \in \arg\max_a A^\pi_{r,k}(s,a)$. For continuous actions, this maximization can be approximated by training an actor (which will be frozen when deploy) while keeping the SUAF fixed. A soft variant can replace the argmax with a Boltzmann update as $\pi^+_r(a|s) \propto \pi(a|s)\exp(\tau A^\pi_{r,k}(s,a))$. Equivalently, π^+_r solves the KL-regularized policy improvement problem $\max_{\pi'} \mathbb{E}_{\pi'}[A^\pi_{r,k}(s,a)] - \tau^{-1}\mathcal{D}_{\mathrm{KL}}(\pi'\|\pi)$. As $\tau \to \infty$, this recovers the greedy update and as $\tau \to 0$, it recovers π. For state-only rewards, $C_\pi r = 0$, so an actor conditioned on $\beta(r)$ can be pretrained across coefficient vectors and reused for new rewards without further parameter updates. With exact full-rank SUAF and exact maximization, π^+_r is greedy with respect to Q^π_r and hence satisfies $V^{\pi^+_r}_r \ge V^\pi_r$. This provides one policy-improvement step. Note that no training steps when deploy to downstream tasks and hence it forms a zero-shot deployment.

Theorem 4.1 (Boltzmann policy improvement with truncated SUAF)**.** *Let $A^\pi_{r,k}$ be the exact rank-k SUAF advantage approximation, and let $\pi_\tau(a|s) \propto \pi(a|s)e^{\tau A^\pi_{r,k}(s,a)}$ with $\tau > 0$. Define $B_k := \sup_{s,a,b} |A^\pi_{r,k}(s,a) - A^\pi_{r,k}(s,b)|$, $\varepsilon_k(r) = \frac{\gamma}{1-\gamma}\left[\sum_{i>k}(\sigma^{+,\pi}_i)^2\beta_i(r)^2\right]^{1/2}$ and $\Delta_k(r) := \|A^\pi_{r,k}\|^2_{d^\pi} - \|C^\pi r\|_{d^\pi}\varepsilon_k(r)$ If $\Delta_k(r) > 0$, then it guarantees $J_r(\pi_\tau) > J_r(\pi)$ if temperature satisfies $0 < \tau < \frac{4(1-\gamma)^2\Delta_k(r)}{R_{\max}B_k^2}$. For state-only rewards, $C^\pi r = 0$ and $\Delta_k(r) = \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_{i\le k}(\sigma^{+,\pi}_i)^2\beta_i(r)^2$. In this case, if $\Delta_k(r) > 0$, the larger sufficient range*

$$0 < \tau < \frac{4\gamma}{R_{\max}B_k^2}\sum_{i\le k}(\sigma^{+,\pi}_i)^2\beta_i(r)^2$$

also guarantees strict improvement.

Policy Gradient with Trust Region Beyond one-step policy improvement, SUAF can also be reused under local policy changes. Trust-region methods constrain policy updates to prevent destructive changes (Schulman et al., 2015; 2017). We show that the error of reusing SUAF from an old policy is controlled within the trust region.

Consider two policies π and π_θ in the same finite MDP with a common initial distribution and $0 < \gamma < 1$. Assume that d^π and $\nu^{+,\pi}$ have full support. Using the SUAF f^π and g^π learned under π, we approximate the advantage under π' by

$$\widehat{A}_r = C_{\pi_\theta} r + \frac{\gamma}{1-\gamma} \sum_{i=1}^{R} \sigma_i^{+,\pi} \big(C_{\pi_\theta} f_i^\pi\big) \beta_i(r), \quad \widehat{A}_r^k = C_{\pi_\theta} r + \frac{\gamma}{1-\gamma} \sum_{i=1}^{k} \sigma_i^{+,\pi} \big(C_{\pi_\theta} f_i^\pi\big) \beta_i(r) \tag{7}$$

where $C_{\pi_\theta} u(s,a) = u(s,a) - \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[u(s,a)]$ and $\beta_i(r) = \langle g_i^\pi, r\rangle_{\nu^{+,\pi}}$. This approximation retains the old features and reward coefficients, while recentering the action-dependent features under the new policy. The discrepancy from $A_r^{\pi_\theta}$ isolates the error caused by reusing the old policy's future information. We measure policy shift by the KL divergence as:

$$\mathcal{D}_{\mathrm{KL}}(\pi_\theta|\pi) := \mathbb{E}_{d_S^{\pi_\theta}} \left[D_{\mathrm{KL}}\big(\pi_\theta(\cdot|s)\|\pi(\cdot|s)\big)\right], \mathcal{D}_{\mathrm{KL}}(\pi\|\pi_\theta) := \mathbb{E}_{d_S^{\pi}} \left[D_{\mathrm{KL}}\big(\pi(\cdot|s)\|\pi_\theta(\cdot|s)\big)\right], \tag{8}$$

where $d_S^\pi(s) = \sum_a d^\pi(s,a)$ is marginal state distribution[4]. Define the action sensitivity

$$L_{\max} := \sup_{\|r\|_{\nu^{+,\pi}} \le 1} \max\{\max_{s,a,b} |A_r^\pi(s,a) - A_r^\pi(s,b)|, \max_{s,a,b} |A_r^{\pi_\theta}(s,a) - A_r^{\pi_\theta}(s,b)|\}. \tag{9}$$

For the policy gradient $\nabla_\theta \bar{J}_r(\pi_\theta) = (1-\gamma)\nabla_\theta J_r(\pi_\theta)$, we approximate the gradient as $\widehat{G}_r = \mathbb{E}_{d^{\pi_\theta}}[\nabla_\theta \log \pi_\theta \widehat{A}_r]$ and rank-k approximation $\widehat{G}_\theta^k(r) = \mathbb{E}_{d^{\pi_\theta}}[\nabla_\theta \log \pi_\theta \, \widehat{A}_r^k]$.

Theorem 4.2 (Policy improvement using transferred SUAF)**.** *Given the policy π_θ in the trust region of π_{ref}, i.e., $\mathcal{D}(\pi_\theta\|\pi_{ref}) < \delta$,, fix a reward r with $\|r\|_{\nu^{+,\pi_{ref}}} \le 1$ and define $u_r := F_\theta^\dagger \widehat{G}_r, s_r := \|\widehat{G}_r\|_{F_\theta^\dagger}, \varepsilon_\delta := (\frac{\gamma L_{\max}^2}{2(1-\gamma)}\delta)^{1/2}$ with $F_\theta = \mathbb{E}_\theta[\nabla_\theta \log \pi_\theta (\nabla_\theta \log \pi_\theta)^T]$ being the Fisher matrix. Then $\nabla_\theta \bar{J}_r(\pi_\theta)^\top u_r \ge s_r(s_r - \varepsilon_\delta)$. In particular, if $s_r > \varepsilon_\delta$, then u_r is a strict ascent direction, and there exists $\eta_0 > 0$ such that $J_r(\pi_{\theta+\eta u_r}) > J_r(\pi_\theta), 0 < \eta < \eta_0$. Moreover, if $\theta \mapsto \bar{J}_r(\pi_\theta)$ has an L-Lipschitz gradient on a convex neighborhood containing the update segment, strict improvement holds whenever*

$$0 < \eta < \frac{2 s_r (s_r - \varepsilon_\delta)}{L \|u_r\|_2^2}.$$

Theorem 4.2 indicates that, when the estimated gradient signal exceeds the approximation error, the resulting natural-gradient direction is a strict ascent direction, and sufficiently small steps improve policy performance. These results align with the trust-region principle and motivate integrating SUAF as a reusable critic into existing policy optimization frameworks, such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017). After training, SUAF and the reward-conditioned actor are frozen. For a new reward, only the linear coefficients are computed, and the policy is obtained by a single forward pass without any actor update.

Theorem 4.3 (Policy-gradient error with truncated SUAF)**.** *Under the assumption in Theorem 4.2. For $\eta_\delta = \min\{1, \sqrt{\delta/[2(1-\gamma)]}\}$, $B(\delta) = \min\{R_{\max}^2, 1 + R_{\max}^2 C_\gamma \sqrt{\delta}\}$ with $C_\gamma = \sqrt{(4+\gamma^{-1})/(2(1-\gamma)^2)}$,*

$$\sup_{\|r\|_{\nu^{+,\pi_\theta}} \le 1} \left\| \nabla_\theta \bar{J}_r(\pi_\theta) - \widehat{G}_\theta^k(r) \right\|_{F_\theta^\dagger}^2 \le B(\delta) \left[\sqrt{\frac{\gamma \delta L_{\max}^2}{2(1-\gamma)}} + \sqrt{\left(\frac{\gamma}{1-\gamma}\right)^2 \left(\sigma_{k+1}^{+,\pi}\right)^2 + L_{\max}^2 \eta_\delta} \right]^2.$$

Soft-HGR Soft-HGR provides an approach to learning feature functions by replacing hard orthogonality constraints with a covariance penalty (Wang et al., 2019). We adapt this objective to the DCDM and use bidirectional temporal-difference (TD) learning to estimate the long-horizon conditional expectations. The detailed process of algorithm can be found in Appendix E.1. We parameterize two k-dimensional feature functions, f_θ and g_ω, with neural networks parameters θ and ω respectively. enforcing $\mathbb{E}_\pi[f_\theta|s] = 0, \mathbb{E}_{\nu^{+,\pi}}[g_\omega] = 0$. These conditions are implemented by subtracting the corresponding conditional and marginal means. In continuous spaces, the means are approximated using samples. Furthermore, defining $C_f = \mathbb{E}_{d^\pi}[f_\theta f_\theta^{\mathrm{T}}]$ and $C_g = \mathbb{E}_{\nu^{+,\pi}}[g_\omega g_\omega^{\mathrm{T}}]$, the feature-learning objective of soft-HGR is

$$\max_{\theta,\omega} \mathcal{J}(\theta,\omega) = \mathbb{E}_{(s,a)\sim d^\pi} \mathbb{E}_{(s^+,a^+)\sim M^{+,\pi}(\cdot|s,a)}[f_\theta^{\mathrm{T}}(s,a) g_\omega(s^+,a^+)] - \frac{1}{2}\operatorname{tr}(C_f C_g). \tag{10}$$

[4] $\mathcal{D}_{\mathrm{KL}}(\pi\|\pi_\theta)$ and $\mathcal{D}_{\mathrm{KL}}(\pi_\theta\|\pi)$ are usually referred as the forward and reverse KL divergence respectively.

The conditional centering of f_θ removes dependence explained solely by the current state, so the objective targets the action-dependent successor residual represented by $\widetilde{\boldsymbol{B}}^{+,\pi}$. Soft-HGR learns low-rank factors rather than individually normalized singular functions. Weighted whitening is then applied to obtain the canonical feature pairs and estimated singular values using sample moments.

Bidirectional conditional prediction. The original soft-HGR method requires full trajectories, which can be a burden. We instead introduce bidirectional Bellman conditional prediction. The correlation term in equation 10 admits two equivalent expressions:

$$\mathbb{E}_{(s,a)\sim d^\pi}\mathbb{E}_{(s^+,a^+)\sim M^{+,\pi}(\cdot|s,a)}[f_\theta^{\mathrm{T}}(s,a)g_\omega(s^+,a^+)] = \mathbb{E}_{d^\pi}[f_\theta^{\mathrm{T}}h] = \mathbb{E}_{\nu^{+,\pi}}[b^{\mathrm{T}}g_\omega], \tag{11}$$

where $h(s,a) = \mathbb{E}[g_\omega(s^+,a^+)|s,a]$ and $b(s^+,a^+) = \mathbb{E}[f_\theta(s,a)|s^+,a^+]$. We estimate these quantities with auxiliary prediction networks. For one-step transitions $(s,a) \sim d^\pi$ and $(s',a') \sim P_\pi(\cdot|s,a)$, the forward predictor and backward recursion satisfy

$$h(s,a) = \mathbb{E}_{P_\pi}[(1-\gamma)g_\omega(s',a') + \gamma h(s',a') \mid s,a],$$
$$v(s',a') = w_1(s',a')\,\mathbb{E}[(1-\gamma)f_\theta(s,a) + \gamma v(s,a) \mid s',a'],$$

where $\nu^{+,\pi}(s,a)b(s,a) = d^\pi(s,a)v(s,a)$ and $w_1(s,a) = \nu_1^\pi(s,a)/d^\pi(s,a)$. Accordingly, networks h_ξ and v_η are trained with TD objectives

$$\mathcal{L}_{\mathrm{F}}(\xi) = \mathbb{E}_{(s,a)\sim d^\pi}\mathbb{E}_{(s',a')\sim P_\pi}\left[\left\|h_\xi(s,a) - \mathrm{sg}\big[(1-\gamma)g_\omega(s',a') + \gamma h_{\bar{\xi}}(s',a')\big]\right\|_2^2\right], \tag{12}$$

$$\mathcal{L}_{\mathrm{B}}(\eta) = \mathbb{E}_{(s',a')\sim\nu_1^\pi}\mathbb{E}_{(s,a)\sim p_\pi}\left[\left\|v_\eta(s',a') - \mathrm{sg}\big[w_1(s',a')\big((1-\gamma)f_\theta(s,a) + \gamma v_{\bar{\eta}}(s,a)\big)\big]\right\|_2^2\right],$$

where sg denotes stop-gradient and $\bar{\xi}, \bar{\eta}$ are target-network parameters. The density ratio w_1 can be eliminated by a surrogate loss $\widehat{\mathcal{L}}_{\mathrm{B}}(\eta)$ sharing the same minimum as $\mathcal{L}_{\mathrm{B}}(\eta)$. We defer its derivation to Appendix E.1.

Alternating soft-HGR updates. With the prediction networks held fixed, we alternate between minimizing

$$\mathcal{L}_f(\theta) = \mathbb{E}_{d^\pi}\left[-f_\theta^{\mathrm{T}}\,\mathrm{sg}[h_\xi] + \frac{1}{2}f_\theta^{\mathrm{T}}\,\mathrm{sg}[C_g]f_\theta\right], \mathcal{L}_g(\omega) = \mathbb{E}_{\nu^{+,\pi}}\left[-g_\omega^{\mathrm{T}}\,\mathrm{sg}[\widehat{b}_\theta] + \frac{1}{2}g_\omega^{\mathrm{T}}\,\mathrm{sg}[C_f]g_\omega\right]. \tag{13}$$

With exact conditional predictors and population moments, these updates are precisely block gradient ascent on equation 10. The forward and backward predictors, h_ξ and v_η are held fixed during each feature update. With exact conditional predictions and population moments, these updates recover the corresponding soft-HGR gradients without differentiating through the temporal-difference (TD) solutions. In practice, prediction and moment-estimation errors introduce errors into these gradient estimates. Expectations are estimated using samples. In particular, $\nu^{+,\pi}$ samples are obtained by selecting an anchor from d^π and a future lag L with $\Pr(L=\ell) = (1-\gamma)\gamma^{\ell-1}$.

5 Experiments

5.1 Experimental Setup

Discrete Finite Environments We evaluate on Grid World (Chevalier-Boisvert et al., 2023), Four-room (Sutton et al., 1999), and Taxi (Dietterich, 2000) to cover basic navigation, narrow passages with long-range credit assignment, and compositional tasks (navigation, pickup, delivery), respectively. These domains vary in state-space size and temporal structure. For each, we partition reward tasks into in-distribution and out-of-distribution sets while fixing dynamics. Representations are learned from reward-free transitions under a behavior policy and rewards are used only for policy extraction. Since states and transitions are enumerable, we compute the optimal return $J(\pi^*)$ via dynamic programming.

Continuous State and Action Environments. We further use four state-based continuous-control domains: walker and cheetah from ExORL (Yarats et al., 2022), and antmaze-large-navigate (denoted as antmaze-large for short) and cube-single-play from OGBench (Park et al., 2025). These jointly cover locomotion under varied rewards, goal navigation, and robotic pick-and-place. Representations are trained on reward-free offline transitions and reused across downstream tasks. As tabular dynamic programming is infeasible, we follow benchmark metrics: undiscounted episodic return for ExORL and task success rate for OGBench.

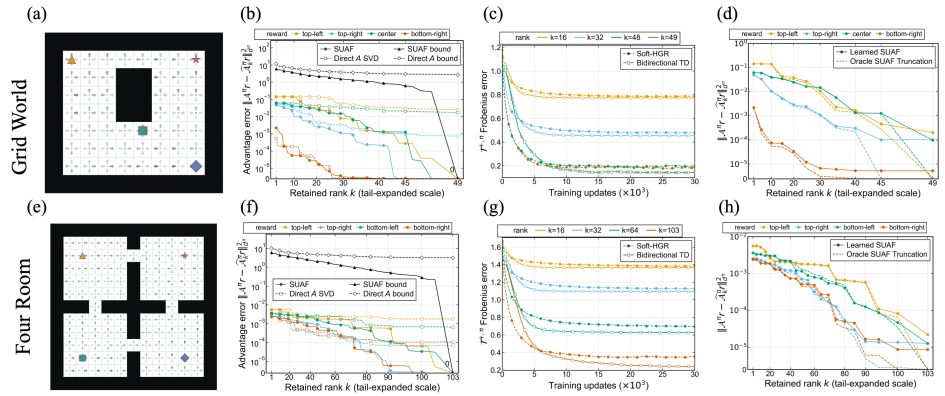


Figure 1: **Spectral approximation and learning of SUAF in Grid World (top) and Four-room (bottom). (a, e)** Environments with reward locations. Background arrows indicate the policy. **(b, f)** Advantage approximation errors for SUAF truncation and direct SVD truncation, together with theoretical worst-case error bounds. **(c, g)** Weighted Frobenius errors for approximating $\mathcal{T}^{+,\pi}$ using Soft-HGR and Bidirectional TD at different ranks. **(d, h)** Reward-specific advantage approximation errors of SUAF learned using Bidirectional TD, compared with exact SUAF truncation computed from the exact operator. Rank axes in (b, d, f, h) use a nonlinear scale that expands the spectral tail.

Table 1: **Zero-shot evaluation on finite state environments.** The performance is measured using the improvement rate defined as $\frac{J(\pi)-J(\pi_\beta)}{J(\pi^*)-J(\pi_\beta)}$, where π^* is the optimal policy, π is the generated policy and π_β is the sample policy. Each result is measured as *mean* with *95% confidence interval* and reports with 5 random sample policies with 5 seeds for each policy.

Domain (Task Num.)	Laplacian	BYOL-γ	TD-JEPA	ICVF	HILP	FB	One-Step FB	Ours
In-Distribution Tasks								
Grid World (4)	$0.006_{\pm 0.030}$	$0.690_{\pm 0.113}$	$\mathbf{0.993}_{\pm 0.015}$	$0.804_{\pm 0.014}$	$0.770_{\pm 0.040}$	$\underline{0.929_{\pm 0.048}}$	$0.599_{\pm 0.036}$	$0.768_{\pm 0.040}$
Four-room (4)	$0.000_{\pm 0.007}$	$0.042_{\pm 0.016}$	$\mathbf{0.900}_{\pm 0.063}$	$0.293_{\pm 0.040}$	$0.450_{\pm 0.167}$	$\underline{0.773_{\pm 0.110}}$	$0.175_{\pm 0.102}$	$0.510_{\pm 0.008}$
Taxi (4)	$0.029_{\pm 0.256}$	$0.064_{\pm 0.009}$	$-0.037_{\pm 0.085}$	$\underline{0.071_{\pm 0.016}}$	$0.042_{\pm 0.113}$	$-0.226_{\pm 0.015}$	$0.067_{\pm 0.040}$	$\mathbf{0.264}_{\pm 0.018}$
OOD Tasks								
Grid World (2)	$0.061_{\pm 0.059}$	$0.443_{\pm 0.055}$	$0.446_{\pm 0.061}$	$0.437_{\pm 0.021}$	$0.405_{\pm 0.044}$	$\underline{0.474_{\pm 0.087}}$	$0.424_{\pm 0.038}$	$\mathbf{0.873}_{\pm 0.016}$
Four-room (2)	$0.070_{\pm 0.028}$	$0.169_{\pm 0.103}$	$0.347_{\pm 0.020}$	$0.283_{\pm 0.085}$	$0.150_{\pm 0.082}$	$\underline{0.385_{\pm 0.073}}$	$0.180_{\pm 0.012}$	$\mathbf{0.636}_{\pm 0.009}$
Taxi (3)	$-0.014_{\pm 0.111}$	$-0.027_{\pm 0.002}$	$0.009_{\pm 0.025}$	$-0.026_{\pm 0.015}$	$-0.021_{\pm 0.048}$	$\underline{0.085_{\pm 0.013}}$	$-0.016_{\pm 0.019}$	$\mathbf{0.300}_{\pm 0.010}$

Baselines For baselines, we compare against methods that are commonly used and are the state of the art in reward-independent features, including Laplacian features (Wu et al., 2019), BYOL-γ (Lawson et al., 2026), TD-JEPA (Bagatella et al., 2026), ICVF (Ghosh et al., 2023), HILP (Park et al., 2024), Forward-Backward (Touati et al., 2023) and One-Step FB (Zheng et al., 2026).

5.2 Results and Analysis

Spectral approximation and learning of SUAF. Figure 1 evaluates SUAF in Grid World and Four-room. SUAF represents the reward-to-advantage map by computing the immediate action contrast $C_\pi r$ exactly and using learned spectral features only for the delayed, action-dependent component. This separation matters because directly compressing the full advantage operator incurs an immediate-term spectral floor when $k < |\mathcal{S}|(|\mathcal{A}| - 1)$. In contrast, the squared error of exact SUAF truncation is bounded, for $\|r\|_{\nu^{+,\pi}} \leq 1$, by $(\gamma/(1-\gamma))^2(\sigma^{+,\pi}_{k+1})^2$, which vanishes when all nonzero delayed modes are retained. Figures 1(b,f) show that the reward-specific errors decrease with retained rank and remain below this worst-case bound. Direct SVD of the full advantage operator has substantially larger errors at the same budget of learned modes. We then test whether the SUAF can be learned from transitions. Across the ranks shown in Figures 1(c,g), Bidirectional TD achieves lower weighted Frobenius error for the delayed operator than Soft-HGR. Finally, Figures 1(d,h) show that the learned SUAF yields decreasing reward-specific advantage errors as rank increases. Its remaining gap from exact SUAF truncation, particularly at higher ranks, reflects feature-learning error in addition to the spectral truncation error.

Table 2: **Zero-shot evaluation on state-based ExORL and OGBench benchmarks.** The full zero-shot evaluation results on 8 ExORL and 10 OGBench tasks. ExORL is measured with returns and OGBench is measured with success rate. The results are reported as *mean* over 8 seeds with *std*.

Domain	Task	Laplacian	BYOL-γ	TD-JEPA	ICVF	HILP	FB	One-Step FB	Ours
ExORL: Locomotion									
walker	overall	228 ± 2	227 ± 2	316 ± 12	$\mathbf{619 \pm 23}$	393 ± 108	400 ± 40	379 ± 26	445 ± 39
	flip	243 ± 5	242 ± 4	381 ± 34	$\mathbf{538 \pm 19}$	332 ± 135	277 ± 100	388 ± 30	357 ± 42
	run	89 ± 1	89 ± 2	112 ± 22	$\mathbf{258 \pm 20}$	136 ± 36	194 ± 20	198 ± 30	210 ± 20
	stand	389 ± 3	387 ± 5	387 ± 11	$\mathbf{858 \pm 31}$	691 ± 126	621 ± 93	524 ± 60	670 ± 100
	walk	192 ± 4	192 ± 6	384 ± 47	$\mathbf{821 \pm 40}$	413 ± 195	506 ± 34	407 ± 63	542 ± 41
cheetah	overall	125 ± 41	127 ± 39	185 ± 7	187 ± 13	116 ± 78	271 ± 46	$\mathbf{378 \pm 56}$	375 ± 21
	run	40 ± 13	42 ± 13	66 ± 3	89 ± 7	38 ± 32	43 ± 36	97 ± 19	$\mathbf{100 \pm 12}$
	run backward	50 ± 17	48 ± 15	69 ± 6	48 ± 9	36 ± 40	125 ± 30	$\mathbf{181 \pm 60}$	159 ± 70
	walk	185 ± 61	199 ± 61	293 ± 15	385 ± 21	195 ± 121	251 ± 166	$\mathbf{424 \pm 44}$	420 ± 32
	walk backward	226 ± 77	220 ± 70	312 ± 26	226 ± 38	194 ± 202	663 ± 143	811 ± 156	$\mathbf{820 \pm 232}$
OGBench: Navigation and manipulation									
antmaze large	overall	9 ± 1	21 ± 2	14 ± 9	23 ± 3	34 ± 2	25 ± 5	30 ± 9	$\mathbf{35 \pm 5}$
	task 1	2 ± 1	6 ± 3	15 ± 11	4 ± 2	13 ± 8	$\mathbf{46 \pm 9}$	21 ± 9	36 ± 8
	task 2	2 ± 1	11 ± 4	1 ± 2	9 ± 3	16 ± 6	2 ± 3	$\mathbf{41 \pm 12}$	24 ± 1
	task 3	29 ± 3	57 ± 8	35 ± 23	67 ± 8	$\mathbf{75 \pm 6}$	31 ± 10	15 ± 4	57 ± 5
	task 4	6 ± 2	14 ± 5	7 ± 7	18 ± 6	27 ± 10	3 ± 2	$\mathbf{33 \pm 15}$	12 ± 6
	task 5	6 ± 3	16 ± 4	11 ± 8	18 ± 4	40 ± 8	$\mathbf{44 \pm 19}$	37 ± 20	$\mathbf{44 \pm 9}$
cube single play	overall	6 ± 2	13 ± 2	8 ± 4	13 ± 2	$\mathbf{30 \pm 8}$	2 ± 1	3 ± 2	21 ± 10
	task 1	5 ± 2	12 ± 4	6 ± 4	13 ± 2	$\mathbf{27 \pm 7}$	1 ± 1	3 ± 4	20 ± 7
	task 2	5 ± 2	13 ± 3	8 ± 6	13 ± 3	$\mathbf{30 \pm 10}$	3 ± 2	4 ± 4	17 ± 12
	task 3	7 ± 3	13 ± 4	14 ± 9	14 ± 4	$\mathbf{23 \pm 13}$	3 ± 3	4 ± 4	18 ± 9
	task 4	6 ± 2	15 ± 6	6 ± 4	11 ± 3	$\mathbf{37 \pm 19}$	2 ± 2	2 ± 2	21 ± 12
	task 5	5 ± 3	11 ± 3	5 ± 4	13 ± 4	$\mathbf{30 \pm 22}$	1 ± 2	1 ± 1	28 ± 20

Zero-shot policy generation ability. (1) **Zero-shot policy extraction in finite-state environments.** Table 1 evaluates the fraction of the improvement from the behavior policy to the optimal policy recovered by each method. Our method achieves the highest mean improvement rate on all three OOD task sets. Performance on in-distribution tasks is more mixed. TD-JEPA and FB outperform SUAF on Grid World and Four-room, whereas SUAF achieves the best result on Taxi. These results support the utility of SUAF for extracting policies across different reward tasks, with particularly favorable performance on the evaluated OOD tasks, while showing that it does not uniformly dominate existing representations. (2) **Evaluation on ExORL and OGBench.** Table 2 examines whether these benefits extend to continuous-state locomotion, navigation, and manipulation tasks. SUAF achieves the second-highest mean return on both ExORL domains. On OGBench, SUAF obtains the highest mean success rate on antmaze large, closely followed by HILP, and ranks second on cube single play, behind HILP. Given the reported variability, the small differences on cheetah and antmaze should be interpreted as comparable performance rather than decisive advantages. Overall, SUAF remains competitive across these domains, while the remaining gaps on walker and cube manipulation indicate room for improvement beyond the finite-state setting.

Why SUAF shows strong OOD generalization? SUAF achieves the highest OOD improvement rates across all three finite-state domains. A potential contributor is its explicit treatment of reward dependence: new rewards enter through linear projections onto shared features, without requiring a task-conditioned advantage predictor. Methods such as FB and TD-JEPA additionally rely on neural predictors conditioned on the task embedding, whose generalization across task embeddings can introduce approximation errors. One-Step FB shares SUAF's task-independent representation and exhibits smaller ID-to-OOD performance changes on the navigation domains, consistent with this interpretation.

6 Conclusion

In this work, we introduced SUAF, a reward-independent representation that preserves the immediate reward contrast and allocates learned capacity to delayed action effects. This decomposition avoids the immediate-term spectral floor of full advantage compression and yields optimal spectral approximations of the delayed operator. We connected representation errors to sufficient conditions for policy improvement and characterized local reuse under policy shifts. Bidirectional TD provides a practical approach to learning the paired features from reward-free data. Experiments support their utility for zero-shot policy extraction across diverse reward tasks. The results highlight the value of separating directly computable reward information from learned temporal effects, while leaving improved estimation of weak spectral modes and reliable reuse over larger policy changes as directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools for (i) language editing, i.e., grammar and clarity revisions of author-written text; (ii) literature search, to help identify potentially relevant prior work; (iii) coding and formatting assistance, including plotting scripts for figures and LaTeX formatting of equations; (iv) drafting sections of the paper, with all AI-generated drafts treated as starting points and subsequently reviewed, edited, and verified by the authors

We have not used generative AI tools to generate, modify, or select experimental data or results, or to fabricate citations.

We have reviewed all AI-assisted work. Every reference surfaced with AI assistance was located and checked against the original source by the authors before inclusion. AI-suggested text edits were reviewed and revised by the authors. AI-assisted plotting code was reviewed by all authors, and each figure was checked against the raw exported results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a complete account of the assumptions and proofs underlying our theoretical results in Appendix D.1 and Appendix D. Detailed experimental settings, including dataset preprocessing, model configurations, training procedures, evaluation protocols, and hyperparameters, are provided in Appendix G. An anonymized implementation of our method, together with the scripts used to reproduce the main experiments, is available at `https://anonymous.4open.science/r/suaf-C01C`. We refer readers to these materials for the exact details needed to reproduce our results.

REFERENCES

Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. Proto successor measure: Representing the behavior space of an RL agent. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=mUDnPzopZF`.

Marco Bagatella, Matteo Pirotta, Ahmed Touati, Alessandro Lazaric, and Andrea Tirinzoni. Td-jepa: Latent-predictive representations for zero-shot reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 36211–36244, 2026.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017a.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017b.

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *arXiv preprint arXiv:2101.07123*, 2021.

Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado Van Hasselt, David Silver, and Tom Schaul. Universal successor features approximators. *arXiv preprint arXiv:1812.07626*, 2018.

David R. Brillinger. Information and information stability of random variables and processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 13(2):134–135, 06 1964. ISSN 0035-9254. doi: 10.2307/2985711. URL `https://doi.org/10.2307/2985711`.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, December 2023.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Chenxiao Gao, Haotian Sun, Na Li, Dale Schuurmans, and Bo Dai. Spectral representation-based reinforcement learning. *arXiv preprint arXiv:2512.15036*, 2025.

Dibya Ghosh, Chethan Anand Bhateja, and Sergey Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, pp. 11321–11339. PMLR, 2023.

Shao-Lun Huang, Anuran Makur, Gregory W. Wornell, and Lizhong Zheng. Universal features for high-dimensional learning and inference. *Foundations and Trends in Communications and Information Theory*, 21(1-2):1–299, 02 2024. ISSN 1567-2190. doi: 10.1561/0100000107. URL https://doi.org/10.1561/0100000107.

Daniel Lawson, Adriana Hugessen, Charlotte Cloutier, Glen Berseth, and Khimya Khetarpal. Self-predictive representations for combinatorial generalization in behavioral cloning. In *International Conference on Learning Representations*, volume 2026, pp. 101559–101586, 2026.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

Yann Ollivier. Which features are best for successor features?, 2025. URL https://arxiv.org/abs/2502.10790.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. *arXiv preprint arXiv:2402.15567*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, volume 2025, pp. 94937–94982, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

John D. Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Dmitry Shribak, Chen-Xiao Gao, Yitong Li, Chenjun Xiao, and Bo Dai. Diffusion spectral representation for reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 110028–110056, 2024.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MYEap_OcQI.

Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang. An efficient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5281–5288, 2019.

Junyi Wu and Dan Li. Factorized spectral representations for reinforcement learning. *arXiv preprint arXiv:2607.13498*, 2026.

Yifan Wu, George Tucker, and Ofir Nachum. The laplacian in rl: Learning representations with efficient approximations. In *International Conference on Learning Representations*, 2019.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025.

Denis Yarats, David Brandfonbrener, Hao Liu, Michael Laskin, Pieter Abbeel, Alessandro Lazaric, and Lerrel Pinto. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. *arXiv preprint arXiv:2201.13425*, 2022.

Chongyi Zheng, Royina Karegoudra Jayanth, and Benjamin Eysenbach. Can we really learn one representation to optimize all rewards? In *Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning*, 2026. URL https://openreview.net/forum?id=ClWj7AfhcY.

A Related Works and Discussion

A.1 Related Works

Successor features and reward factorization. Classical successor features (Barreto et al., 2017a; Borsa et al., 2018) model $r_w = \phi^{\mathrm{T}} w$, yielding $Q^\pi_{r_w} = (\psi^\pi)^{\mathrm{T}} w$, with ψ^π accumulating discounted features. SUAF instead derives g_i from delayed credit, with readouts $\beta_i^{+,\pi}(r) = \langle g_i, r\rangle_{\nu^{+,\pi}}$. The residual $r_\perp = r - \sum_{i=1}^R \beta_i^{+,\pi}(r) g_i$, orthogonal to all R nonzero modes, contributes no delayed advantage, while $C_\pi r$ remains its immediate effect. Truncating to $k < R$ can discard decision-relevant modes, with error $\|A^\pi_r - A^\pi_{r,k}\|^2_{d^\pi} = (\frac{\gamma}{1-\gamma})^2 \sum_{i=k+1}^R (\sigma_i^{+,\pi})^2 [\beta_i^{+,\pi}(r)]^2$. Therefore, SUAF optimizes advantage reconstruction, not reward reconstruction.

Optimal successor features. Ollivier (2025) characterizes the leading right singular functions of $\mathcal{A}_\pi = C_\pi (I - \gamma P_\pi)^{-1}$ through asymptotic downstream performance under strong behavior-cloning regularization, without providing a learning algorithm. If under matching stationary measures and L^2 geometry, the full DCDM has the same right singular subspaces. However, our g_i instead diagonalize $(\mathcal{D}^+_\pi)^* \mathcal{D}^+_\pi$, where $\mathcal{D}^+_\pi = \mathcal{A}_\pi - C_\pi = \gamma C_\pi P_\pi (I - \gamma P_\pi)^{-1}$, and generally differ. In doing this, we are able to get a possibly close to 0 approximation loss to $\mathcal{A}_\pi$, which enables the control of policy gradient error. See detailed discussions in Appendix D.2. Our rank-k optimality is exact for the fixed-policy delayed operator. We jointly estimate paired action-credit and reward features using Soft-HGR (Wang et al., 2019) and bidirectional TD. We emphasize that $C_\pi r = 0$ on the subspace of state-only rewards does not change the full-space operator underlying SUAF. SUAF decomposes the action-centered positive-lag DCDM, whereas the full successor measure, the full advantage operator, and the one-step transition kernel define different spectral problems and generally yield different features.

Proto successor measures and scope. Proto Successor Measures (Agarwal et al., 2025) represent feasible successor measures across policies in a policy-independent affine basis, $M^\pi = b + \sum_\ell w^\pi_\ell \Phi_\ell$. For $0 < \gamma < 1$, $M^\pi = (1-\gamma) I + \gamma M^{+,\pi}$ links PSM to $\widetilde{\boldsymbol{B}}^{+,\pi} = D^{1/2}_{d^\pi} C_\pi M^{+,\pi} D^{-1/2}_{\nu^{+,\pi}}$, where $D_\mu = \mathrm{diag}(\mu)$. It removes action-invariant occupancy and instead encodes rewards through $\beta^{+,\pi}(r)$ at fixed π. Exact PSM determines DCDM, but truncated DCDM cannot generally reconstruct the full successor measure. Policy-dependent centering and marginals prevent automatically inheriting PSM's cross-policy affine structure. Our scope is fixed-policy reward reuse, not policy-independent optimal control. For zero-shot learning with SUAF, please refer to Appendix D.5.

Spectral Reinforcement Learning. Existing spectral representations differ in the operator whose structure they retain (Table 3). Gao et al. (Gao et al., 2025) study spectral representations of the one-step transition operator, while Diff-SR (Shribak et al., 2024) extracts transition-based features through diffusion and energy-based models. FaStR (Wu & Li, 2026) further factorizes the transition kernel across states, actions, and next states. Moving beyond one-step dynamics, One-Step FB (Zheng et al., 2026) represents the successor measure of a fixed behavior policy for downstream policy improvement. Ollivier (Ollivier, 2025) selects successor features using a spectral criterion derived from downstream performance under strong behavior-cloning regularization. SUAF instead applies action centering to the positive-lag successor measure before spectral decomposition. This yields a reward-independent representation of *delayed advantage*, while the immediate reward contrast is computed exactly; the rank budget is therefore devoted to the component that must be approximated.

A.2 Difference and relations between SUAF and existing operators

Figure 2 compares the population feature subspaces induced by SUAF, the full advantage operator, One-Step FB, Ollivier, and PSM. All operators are constructed exactly under the same reference policy, so the differences do not arise from optimization noise.

The left column shows the singular spectra. For SUAF, we plot the singular values of the delayed contribution $D^\pi := \frac{\gamma}{1-\gamma} \mathcal{T}^{+,\pi}$, whereas the full-advantage curve is obtained from $A^\pi = C_\pi + D^\pi$. The dotted horizontal line marks the unit spectral floor contributed by the immediate action-centering operator C_π. In GridWorld and Four Rooms, it recovers

$$\sigma_i(A^\pi) = \sqrt{1 + \sigma_i(D^\pi)^2}.$$

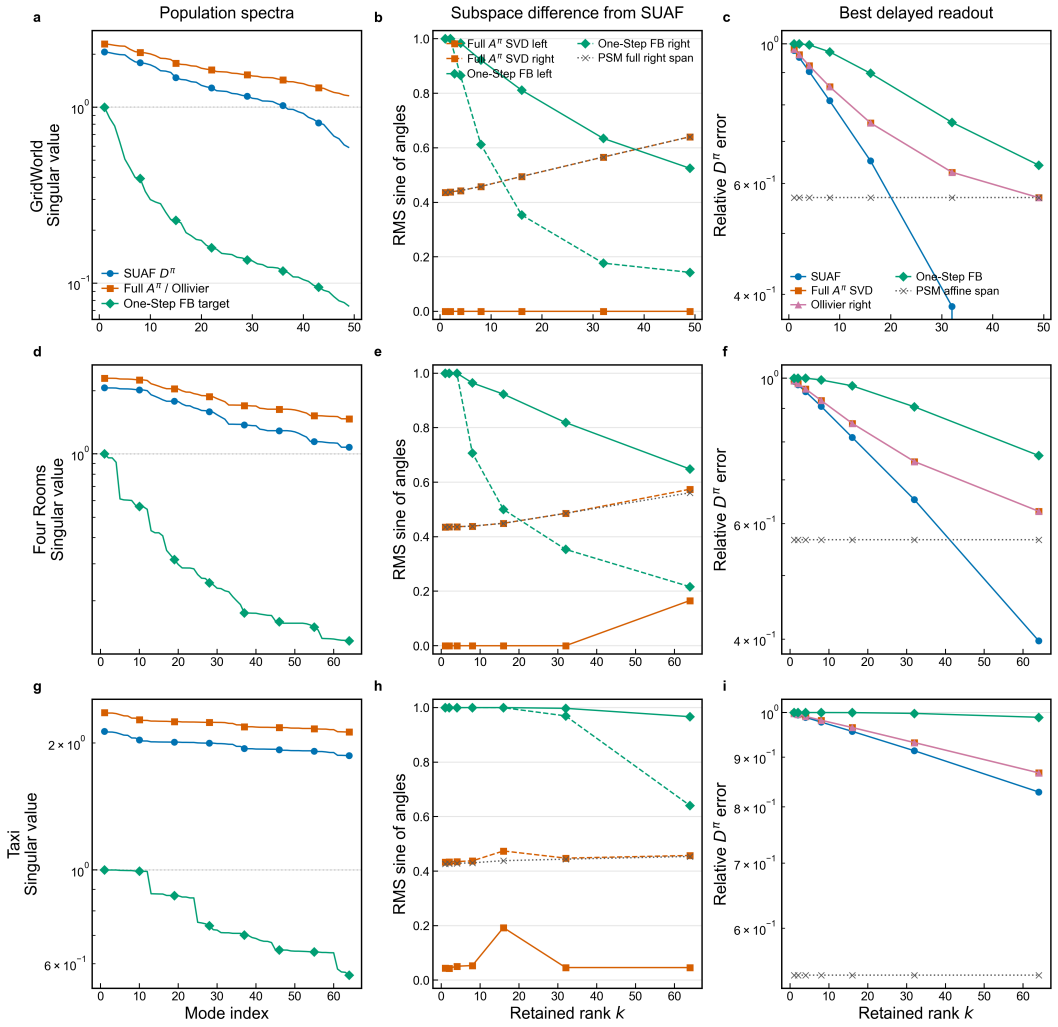


Figure 2: **The population feature subspaces induced by SUAF, the full advantage operator, One-Step FB, Ollivier, and PSM.** Since all the space are solved explicitly, hence no seeds are needed here.

Table 3: **Comparison of the operators underlying related representations.** Transition-based methods represent one-step dynamics. SUAF decomposes the action-centered positive-lag successor operator with immediate reward.

Representation	Represented or decomposed object
Gao et al. (Gao et al., 2025)	One-step transition kernel $P(s'\|s,a)$; spectral factorization.
Diff-SR (Shribak et al., 2024)	One-step $P(s'\|s,a)$; diffusion/energy-based spectral features.
FaStR (Wu & Li, 2026)	Transition tensor: $P(s'\|s,a) \approx \sum_i \phi_i^S(s)\phi_i^A(a)m_i(s')$ (CP factorization).
One-Step FB (Zheng et al., 2026)	Successor-measure ratio $M^{\pi_\beta}(y\|x)/\rho(y)$ for a fixed behavior policy.
Ollivier (Ollivier, 2025)	Spectral criterion based on downstream performance; involves the full advantage operator $C^\pi(I-\gamma P^\pi)^{-1}$
SUAF (ours)	$\mathrm{svd}\Big(D_{d^\pi}^{1/2} C^\pi M^{+,\pi} D_{\nu^{+,\pi}}^{-1/2}\Big)$; $A_r^\pi = C^\pi r + \frac{\gamma}{1-\gamma} C^\pi M^{+,\pi} r$.

Consequently, the spectrum of the full advantage cannot decay below one even when the delayed singular values are small. Directly truncating A^π therefore spends rank on the analytically known immediate term, whereas SUAF assigns its entire rank budget to delayed credit assignment. The One-Step FB spectrum is that of the uncentered positive-lag successor kernel; its absolute scale depends on the successor-measure normalization, so this curve is primarily used to compare spectral decay rather than magnitude.

The middle column measures the difference between feature subspaces. Let U_k^D, V_k^D denote the leading SUAF left and right singular vectors and let U_k^m, V_k^m denote those of method m. We compute the principal angles through

$$\cos\theta_i^L = \sigma_i\big((U_k^D)^\top U_k^m\big), \qquad \cos\theta_i^R = \sigma_i\big((V_k^D)^\top V_k^m\big),$$

and report

$$\text{RMS-sin} = \left(\frac{1}{k}\sum_{i=1}^{k}\sin^2\theta_i\right)^{1/2}.$$

A value of zero means that the two subspaces coincide, whereas a value near one means that they are nearly orthogonal. Although the left singular subspaces of D^π and A^π coincide in the stationary GridWorld and Four Rooms constructions, their right subspaces differ substantially. At $k = 8$, the right-subspace RMS sine is 0.458 in GridWorld, 0.439 in Four Rooms, and 0.438 in Taxi. Since the features proposed by Ollivier are the right singular functions of A^π, the Ollivier and full-A^π right subspaces are identical by construction.

The discrepancy with One-Step FB is larger. One-Step FB factorizes an uncentered successor measure and therefore retains future information that is predictable from the current state alone. SUAF instead subtracts this state-predictable component and retains only the incremental effect of the current action. Accordingly, at $k = 8$, the right-subspace RMS sine between SUAF and One-Step FB is 0.612, 0.707, and 1.000 in the three environments, respectively. These principal angles provide direct spectral evidence that SUAF does not merely learn a rotated version of the One-Step FB representation.

The right column evaluates how well each subspace reconstructs the delayed operator. For a method providing left and right subspaces, we report

$$\mathrm{Err}_D^m(k) = \frac{\big\|D^\pi - P_{U_k^m} D^\pi P_{V_k^m}\big\|_{\mathrm{F}}}{\|D^\pi\|_{\mathrm{F}}},$$

where $P_U = UU^\top$. For Ollivier, only the right subspace is prescribed, so we use the best unrestricted left readout,

$$\mathrm{Err}_D^{\mathrm{Oll}}(k) = \frac{\|D^\pi - D^\pi P_{V_k^A}\|_{\mathrm{F}}}{\|D^\pi\|_{\mathrm{F}}}.$$

For SUAF, this error reduces to

$$\mathrm{Err}_D^{\mathrm{SUAF}}(k) = \left(\frac{\sum_{i>k}\sigma_i(D^\pi)^2}{\sum_i \sigma_i(D^\pi)^2}\right)^{1/2},$$

which is the Eckart-Young optimal rank-k error. SUAF consequently achieves the lowest delayed-operator error among all rank-matched methods at every evaluated rank. For example, at $k = 8$, the errors of SUAF, full-A^π, and One-Step FB are respectively $0.812/0.855/0.971$ in GridWorld, $0.906/0.925/0.994$ in Four Rooms, and $0.978/0.982/1.000$ in Taxi.

The PSM curve is not a rank-matched baseline. It projects onto the complete policy-independent affine span, whose dimensions are 200, 416, and 2600 in the three environments. It is therefore shown only as a containment reference and should not be interpreted as outperforming a rank-k representation.

A.3 Properties of DCDM

A.3.1 Rank bound of the DCDM

For a finite MDP, the positive-lag future advantage operator satisfies $R := \mathrm{rank}(\mathcal{T}^{+,\pi}) \le |\mathcal{S}| - 1$, regardless of the number of actions. To see this, define the policy-averaged reward $\bar{r}_\pi(s) := \sum_a \pi(a \mid s) r(s,a)$. Since all future actions are sampled from π, the future term depends on r only through $\bar{r}_\pi$. More explicitly, let $(Tv)(s,a) := \sum_{s'} T(s' \mid s,a) v(s')$ and let P_S^π be the state transition operator under π. Then

$$\mathcal{T}^{+,\pi} r = (1-\gamma) C_\pi T (I - \gamma P_S^\pi)^{-1} \bar{r}_\pi.$$

Thus, $\mathcal{T}^{+,\pi}$ factors through the $|\mathcal{S}|$-dimensional space of state functions. Moreover, a constant $\bar{r}_\pi$ produces no action contrast, since $(I - \gamma P_S^\pi)^{-1}\mathbf{1} = (1-\gamma)^{-1}\mathbf{1}$ and $C_\pi T \mathbf{1} = 0$. Consequently, at most $|\mathcal{S}| - 1$ nonzero SUAF pairs are needed to represent the future advantage term exactly for every reward.

A.3.2 Information Background of DCDM

We show that the property $\left\| \widetilde{\boldsymbol{B}}^{+,\pi} \right\|_{\mathrm{F}}^2 \approx 2I(A; S^+, A^+ | S)$ holds under the weak dependency condition. For simplicity, denote $\delta(s^+, a^+, s, a) = \frac{M^{+,\pi}(s^+, a^+ | s, a)}{\nu^{+,\pi}(s^+, a^+)} - 1$, i.e., $M^{+,\pi}(s^+, a^+ \mid s, a) = \nu^{+,\pi}(s^+, a^+)(1 + \delta(s^+, a^+, s, a))$. It follows that

$$\bar{M}^{+,\pi}(s^+, a^+ \mid s) = \mathbb{E}_{a' \sim \pi(\cdot|s)} \left[M^{+,\pi}(s^+, a^+ | s, a') \right] = \nu^{+,\pi}(s^+, a^+)(1 + \bar{\delta}(s^+, a^+, s)),$$

where $\bar{\delta}(s^+, a^+, s) = \sum_{a'} \pi(a' \mid s) \delta(s^+, a^+, s, a')$. We also have that

$$\sum_{s^+, a^+, s, a} d^\pi(s,a) \nu^{+,\pi}(s^+, a^+) \delta(s^+, a^+, s, a) = 0, \tag{14}$$

$$\sum_{s^+, a^+, s} d^\pi(s,a) \nu^{+,\pi}(s^+, a^+) \delta(s^+, a^+, s) = 0. \tag{15}$$

Assume that $|\delta(s^+, a^+, s, a)| < \epsilon, \forall s^+, a^+, s, a$ follows from the weak dependency condition, then $|\bar{\delta}(s^+, a^+, s)| < \epsilon, \forall s^+, a^+, s$. Computation gives that $I(A; S^+, A^+ \mid S) = \frac{1}{2} \left\| \widetilde{\boldsymbol{B}}^{+,\pi} \right\|_{\mathrm{F}}^2 + O(\epsilon^3)$

as follows:

$$
\begin{aligned}
& I(A; S^+, A^+ \mid S) \\
&= \sum_{s^+,a^+,s,a} d^\pi(s,a) M^{+,\pi}(s^+, a^+ \mid s, a) \ln \frac{M^{+,\pi}(s^+, a^+ \mid s, a)}{\bar{M}^{+,\pi}(s^+, a^+ \mid s)} && (16) \\
&= \sum_{s^+,a^+,s,a} d^\pi(s,a) \nu^{+,\pi}(s^+, a^+)(1 + \delta(s^+, a^+, s, a)) \ln \frac{1 + \delta(s^+, a^+, s, a)}{1 + \bar{\delta}(s^+, a^+, s)} && (17) \\
&= \sum_{s^+,a^+,s,a} d^\pi(s,a) \nu^{+,\pi}(s^+, a^+) \cdot \\
&\qquad \left(\delta(s^+, a^+, s, a) - \bar{\delta}(s^+, a^+, s) + \frac{1}{2}(\delta(s^+, a^+, s, a) - \bar{\delta}(s^+, a^+, s))^2 + O\left(\epsilon^3\right) \right) && (18) \\
&= \sum_{s^+,a^+,s,a} d^\pi(s,a) \nu^{+,\pi}(s^+, a^+) \cdot \frac{1}{2}(\delta(s^+, a^+, s, a) - \bar{\delta}(s^+, a^+, s))^2 + O\left(\epsilon^3\right) && (19) \\
&= \frac{1}{2} \sum_{s^+,a^+,s,a} d^\pi(s,a) \frac{(M^{+\pi}(s^+, a^+ \mid s, a) - \bar{M}^{+\pi}(s^+, a^+ \mid s))^2}{\nu^{+,\pi}(s^+, a^+)} + O\left(\epsilon^3\right) && (20) \\
&= \frac{1}{2} \left\| \widetilde{\boldsymbol{B}}^{+,\pi} \right\|_{\mathrm{F}}^2 + O(\epsilon^3), && (21)
\end{aligned}
$$

where from to equation 18 to equation 19 we have used equation 14 and equation 15.

A.4 Properties of the proposed SUAF

The proposed SUAF satisfies the following properties: (1) **Weighted orthonormality.** For all $i, j \leq R$, $\mathbb{E}_{d^\pi}[f_i f_j] = \mathbb{E}_{\nu^{+,\pi}}[g_i g_j] = \mathbf{1}(i = j)$; (2) **Centering.** $\mathbb{E}_{\pi(\cdot|s)}[f_i] = \mathbb{E}_{\nu^{+,\pi}}[g] = 0$; (3) **Diagonal canonical dependence.** $\mathbb{E}_{(s,a)\sim d^\pi, (s^+,a^+)\sim M^{+,\pi}(\cdot|s,a)}\left[f_i(s,a) g_j(s^+,a^+)\right] = \sigma_i^{+,\pi} \mathbf{1}(i = j)$; (4) **Variational optimality.** The first feature pair solves $\sigma_1^{+,\pi} = \max_{f,g} \mathbb{E}\left[f(s,a) g(s^+,a^+)\right]$, subject to $\mathbb{E}_\pi[f(s,a)|s] = \mathbb{E}_{\nu^{+,\pi}}[g] = 0, \|f\|_{d^\pi} = \|g\|_{\nu^{+,\pi}} = 1$. Each subsequent pair (f_i, g_i) solves the same problem subject additionally to weighted orthogonality with all preceding feature pairs. Furthermore, since both feature families are centered and normalized, $\sigma_i^{+,\pi}$ is the correlation carried by the i-th feature pair, and hence $0 \leq \sigma_i^{+,\pi} \leq 1$.

A.5 Reward Shaping in the Sense of SUAF

The accuracy for an individual reward depends on the weighted spectral coefficients:

$$
\left\| A_r^\pi - A_{r,k}^\pi \right\|_{d^\pi}^2 = \left(\frac{\gamma}{1-\gamma} \right)^2 \sum_{i=k+1}^{R} \left(\sigma_i^{+,\pi} \right)^2 \left| \langle g_i, r \rangle_{\nu^{+,\pi}} \right|^2 . \tag{22}
$$

Consequently, rewards with substantial energy in the discarded modes can be poorly approximated. To address this problem, we consider potential-based reward shaping that redistributes credit between the exact immediate term and the approximated future term to reduce this error while preserving the exact advantage function. For a bounded state potential Φ, define

$$
r_\Phi(s, a, s') = r(s, a) + \gamma \Phi(s') - \Phi(s), \tag{23}
$$

In the discounted infinite-horizon setting, this transformation gives $Q_{r_\Phi}^\pi(s,a) = Q_r^\pi(s,a) - \Phi(s)$ and $V_{r_\Phi}^\pi(s) = V_r^\pi(s) - \Phi(s)$, hence $A_{r_\Phi}^\pi = A_r^\pi$. To apply the state-action reward representation, let $\bar{r}_\Phi = r + \gamma T\Phi - \Phi$, where $(T\Phi)(s,a) = \mathbb{E}[\Phi(s') \mid s, a]$. Let Π_k denote the $L^2(d^\pi)$-orthogonal projection onto $\mathrm{span}\{f_1, \ldots, f_k\}$.

Theorem A.1 (Potential-dependent approximation error)**.** *Let $0 < \gamma < 1$ and let r be bounded. For $i = k+1, \ldots, R$, define the tail vectors $b_i = \sigma_i \langle g_i, r \rangle_\nu$, and $z_i(\Phi) = \sigma_i \langle g_i, \gamma T\Phi - \Phi \rangle_\nu$. Let $c = \gamma/(1-\gamma)$. The approximation $A_{\bar{r}_\Phi,k}^\pi = C_\pi \bar{r}_\Phi + c\Pi_k \mathcal{T}^{+,\pi} \bar{r}_\Phi$ satisfies*

$$
E_k(\Phi_w; r) := \|A_r^\pi - A_{\bar{r}_\Phi,k}^\pi\|_{d^\pi}^2 = c^2 \|b + z(\Phi)\|_2^2 = \gamma^2 \|(I - \Pi_k) C_\pi T(V_r^\pi - \Phi)\|_{d^\pi}^2 . \tag{24}
$$

Consequently, shaping does not increase the error if and only if $2\langle b, z(\Phi)\rangle + \|z(\Phi)\|_2^2 \leq 0$.

Table 4: **Zero-shot evaluation on finite state environments.** The performance is measured using the improvement rate defined as $\frac{J(\pi)-J(\pi_\beta)}{J(\pi^*)-J(\pi_\beta)}$, where π^* is the optimal policy solved by linear program, π is the generated policy and π_β is the sample policy. Each result is measured as *mean* with *95% confidence interval* and reports with 5 random sample policies with 5 seeds for each policy.

Domain	Task	Laplacian	BYOL-γ	TD-JEPA	ICVF	HILP	FB	One-Step FB	Ours
In-Distribution Tasks									
Grid World	**overall**	$0.006_{\pm 0.030}$	$0.690_{\pm 0.113}$	$\mathbf{0.993}_{\pm 0.015}$	$0.804_{\pm 0.014}$	$0.770_{\pm 0.040}$	$0.929_{\pm 0.048}$	$0.599_{\pm 0.036}$	$0.768_{\pm 0.040}$
	top-left	$-0.016_{\pm 0.124}$	$0.967_{\pm 0.033}$	$0.984_{\pm 0.045}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$0.979_{\pm 0.038}$	$0.767_{\pm 0.126}$	$0.893_{\pm 0.057}$
	top-right	$-0.002_{\pm 0.052}$	$0.060_{\pm 0.143}$	$0.998_{\pm 0.006}$	$\mathbf{1.000}_{\pm 0.000}$	$0.931_{\pm 0.186}$	$0.868_{\pm 0.095}$	$0.517_{\pm 0.033}$	$0.695_{\pm 0.023}$
	center	$-0.001_{\pm 0.069}$	$0.734_{\pm 0.438}$	$0.994_{\pm 0.016}$	$0.997_{\pm 0.009}$	$\mathbf{1.000}_{\pm 0.000}$	$0.895_{\pm 0.153}$	$0.760_{\pm 0.051}$	$0.849_{\pm 0.040}$
	bottom-right	$0.042_{\pm 0.103}$	$\mathbf{1.000}_{\pm 0.000}$	$0.996_{\pm 0.012}$	$0.218_{\pm 0.060}$	$0.150_{\pm 0.030}$	$0.974_{\pm 0.045}$	$0.355_{\pm 0.063}$	$0.635_{\pm 0.178}$
Four-room	**overall**	$0.000_{\pm 0.007}$	$0.042_{\pm 0.016}$	$\mathbf{0.900}_{\pm 0.063}$	$0.293_{\pm 0.040}$	$0.450_{\pm 0.167}$	$0.773_{\pm 0.110}$	$0.175_{\pm 0.102}$	$0.510_{\pm 0.008}$
	top-left	$-0.003_{\pm 0.010}$	$0.042_{\pm 0.028}$	$\mathbf{0.919}_{\pm 0.045}$	$0.278_{\pm 0.206}$	$0.642_{\pm 0.207}$	$0.508_{\pm 0.006}$	$0.022_{\pm 0.003}$	$0.671_{\pm 0.031}$
	top-right	$-0.001_{\pm 0.010}$	$0.056_{\pm 0.012}$	$\mathbf{0.962}_{\pm 0.037}$	$0.261_{\pm 0.096}$	$0.236_{\pm 0.155}$	$0.961_{\pm 0.042}$	$0.071_{\pm 0.149}$	$0.415_{\pm 0.000}$
	bottom-left	$0.003_{\pm 0.016}$	$0.054_{\pm 0.072}$	$0.825_{\pm 0.220}$	$0.348_{\pm 0.150}$	$0.582_{\pm 0.468}$	$\mathbf{0.849}_{\pm 0.212}$	$0.250_{\pm 0.260}$	$0.528_{\pm 0.011}$
	bottom-right	$0.000_{\pm 0.010}$	$0.018_{\pm 0.037}$	$\mathbf{0.894}_{\pm 0.040}$	$0.286_{\pm 0.146}$	$0.340_{\pm 0.227}$	$0.772_{\pm 0.324}$	$0.357_{\pm 0.063}$	$0.425_{\pm 0.000}$
Taxi	**overall**	$0.029_{\pm 0.256}$	$0.064_{\pm 0.009}$	$-0.037_{\pm 0.085}$	$0.071_{\pm 0.016}$	$0.042_{\pm 0.113}$	$-0.226_{\pm 0.015}$	$0.067_{\pm 0.040}$	$\mathbf{0.264}_{\pm 0.018}$
	delivery[1]	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
	fast delivery[2]	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
	safe delivery	$0.128_{\pm 1.033}$	$0.271_{\pm 0.035}$	$-0.158_{\pm 0.323}$	$0.297_{\pm 0.067}$	$0.179_{\pm 0.459}$	$-0.904_{\pm 0.051}$	$0.262_{\pm 0.158}$	$\mathbf{0.453}_{\pm 0.000}$
	staged service	$-0.012_{\pm 0.009}$	$-0.015_{\pm 0.000}$	$0.010_{\pm 0.033}$	$-0.012_{\pm 0.004}$	$-0.012_{\pm 0.008}$	$0.000_{\pm 0.007}$	$0.007_{\pm 0.016}$	$\mathbf{0.605}_{\pm 0.073}$
OOD Tasks									
Grid World	**overall**	$0.061_{\pm 0.059}$	$0.443_{\pm 0.055}$	$0.446_{\pm 0.061}$	$0.437_{\pm 0.021}$	$0.405_{\pm 0.044}$	$0.474_{\pm 0.087}$	$0.424_{\pm 0.038}$	$\mathbf{0.873}_{\pm 0.016}$
	state composition	$0.078_{\pm 0.069}$	$\mathbf{0.923}_{\pm 0.061}$	$0.839_{\pm 0.110}$	$0.811_{\pm 0.043}$	$0.743_{\pm 0.104}$	$0.890_{\pm 0.106}$	$0.777_{\pm 0.044}$	$0.771_{\pm 0.040}$
	state-action circulation	$0.045_{\pm 0.074}$	$-0.036_{\pm 0.060}$	$0.053_{\pm 0.031}$	$0.063_{\pm 0.020}$	$0.067_{\pm 0.028}$	$0.058_{\pm 0.073}$	$0.072_{\pm 0.041}$	$\mathbf{0.975}_{\pm 0.021}$
Four-room	**overall**	$0.070_{\pm 0.028}$	$0.169_{\pm 0.103}$	$0.347_{\pm 0.020}$	$0.283_{\pm 0.085}$	$0.150_{\pm 0.082}$	$0.385_{\pm 0.073}$	$0.180_{\pm 0.012}$	$\mathbf{0.636}_{\pm 0.009}$
	state composition	$0.068_{\pm 0.029}$	$0.258_{\pm 0.201}$	$0.602_{\pm 0.045}$	$0.509_{\pm 0.209}$	$0.337_{\pm 0.155}$	$\mathbf{0.660}_{\pm 0.117}$	$0.321_{\pm 0.026}$	$0.640_{\pm 0.017}$
	state-action circulation	$0.071_{\pm 0.028}$	$0.081_{\pm 0.014}$	$0.092_{\pm 0.065}$	$0.057_{\pm 0.042}$	$-0.037_{\pm 0.049}$	$0.111_{\pm 0.068}$	$0.038_{\pm 0.017}$	$\mathbf{0.632}_{\pm 0.007}$
Taxi	**overall**	$-0.014_{\pm 0.111}$	$-0.027_{\pm 0.002}$	$0.009_{\pm 0.025}$	$-0.026_{\pm 0.015}$	$-0.021_{\pm 0.048}$	$0.085_{\pm 0.013}$	$-0.016_{\pm 0.019}$	$\mathbf{0.300}_{\pm 0.010}$
	pickup priority	$-0.012_{\pm 0.009}$	$-0.015_{\pm 0.000}$	$0.011_{\pm 0.023}$	$-0.012_{\pm 0.002}$	$-0.012_{\pm 0.003}$	$0.025_{\pm 0.025}$	$0.051_{\pm 0.022}$	$\mathbf{0.642}_{\pm 0.029}$
	center hazard	$\mathbf{0.005}_{\pm 0.000}$	$\mathbf{0.005}_{\pm 0.000}$	$-0.025_{\pm 0.041}$	$\mathbf{0.005}_{\pm 0.000}$	$0.004_{\pm 0.004}$	$-0.019_{\pm 0.027}$	$-0.032_{\pm 0.023}$	$\mathbf{0.005}_{\pm 0.000}$
	action fuel	$-0.036_{\pm 0.324}$	$-0.072_{\pm 0.007}$	$0.041_{\pm 0.056}$	$-0.071_{\pm 0.044}$	$-0.054_{\pm 0.140}$	$0.248_{\pm 0.013}$	$-0.066_{\pm 0.045}$	$\mathbf{0.253}_{\pm 0.001}$

[1] For solution to sparse reward, please refer to Appendix B.3 with reward shaping. For fair comparison, we will not show the result here.
[2] Fast delivery is the same with delivery environment.

Therefore, potential quality is determined by the unremained action-credit component of its value error, rather than by ordinary value prediction error alone.

Corollary A.2 (Optimal linear combination of potentials)**.** *Fix bounded state features $h_1, \dots, h_p$ and set $\Phi_w = \sum_{\ell=1}^{p} w_\ell h_\ell$. Let $D = [z(h_1), \dots, z(h_p)]$. A minimum-norm optimal coefficient vector and the resulting error are*

$$w^\star = -D^\dagger b, \quad \min_w E_k(\Phi_w; r) = c^2 \|(I - DD^\dagger)b\|_2^2.$$

In particular, zero error is attainable within this class if and only if $b \in \mathrm{col}(D)$.

Without restricting the potential class, the oracle choice $\Phi = V_r^\pi$ yields zero error, although obtaining this potential requires policy evaluation and will cancel out all the inner product terms $\langle g_i, r_\Phi \rangle$. In contrast, we propose that move the reward to the head of spectrum also help SUAF algorithms estimate advantage better. For a specific example, please refer to B.3.

B Extra Experiments

B.1 Full result

In this subsection, we report our complete results, including the mean and statistical results (e.g., 95% confidence intervals) across all tasks in both discrete and continuous environments. The results are summarized in Table 4 and Table 5.

B.2 One-Step Policy Improvement with SUAF

For reward r and retained rank k, the learned SUAF critic is

$$\widehat{A}^{\pi_0}_{r,k} = C_{\pi_0} r + \frac{\gamma}{1-\gamma} \sum_{i=1}^{k} \widehat{\sigma}_i \widehat{f}_i \langle \widehat{g}_i, r \rangle_{\widehat{\nu}}. \tag{25}$$

Table 5: **Zero-shot evaluation on state-based ExORL and OGBench benchmarks.** The full zero-shot evaluation results on 8 ExORL tasks and 10 OGBench tasks. The results are reported as *mean* over 8 seeds with *std*.

Domain	Task	Laplacian	BYOL-γ	TD-JEPA	ICVF	HILP	FB	One-Step FB	Ours
ExORL: Locomotion									
walker	overall	228 ± 2	227 ± 2	316 ± 12	$\mathbf{619 \pm 23}$	393 ± 108	400 ± 40	379 ± 26	445 ± 39
	flip	243 ± 5	242 ± 4	381 ± 34	$\mathbf{538 \pm 19}$	332 ± 135	277 ± 100	388 ± 30	357 ± 42
	run	89 ± 1	89 ± 2	112 ± 22	$\mathbf{258 \pm 20}$	136 ± 36	194 ± 20	198 ± 30	210 ± 20
	stand	389 ± 3	387 ± 5	387 ± 11	$\mathbf{858 \pm 31}$	691 ± 126	621 ± 93	524 ± 60	670 ± 100
	walk	192 ± 4	192 ± 6	384 ± 47	$\mathbf{821 \pm 40}$	413 ± 195	506 ± 34	407 ± 63	542 ± 41
cheetah	overall	125 ± 41	127 ± 39	185 ± 7	187 ± 13	116 ± 78	271 ± 46	$\mathbf{378 \pm 56}$	375 ± 21
	run	40 ± 13	42 ± 13	66 ± 3	89 ± 7	38 ± 32	43 ± 36	97 ± 19	$\mathbf{100 \pm 12}$
	run backward	50 ± 17	48 ± 15	69 ± 6	48 ± 9	36 ± 40	125 ± 30	$\mathbf{181 \pm 60}$	159 ± 70
	walk	185 ± 61	199 ± 61	293 ± 15	385 ± 21	195 ± 121	251 ± 166	$\mathbf{424 \pm 44}$	420 ± 32
	walk backward	226 ± 77	220 ± 70	312 ± 26	226 ± 38	194 ± 202	663 ± 143	811 ± 156	$\mathbf{820 \pm 232}$
OGBench: Navigation and manipulation									
antmaze large	overall	9 ± 1	21 ± 2	14 ± 9	23 ± 3	34 ± 2	25 ± 5	30 ± 9	$\mathbf{35 \pm 5}$
	task 1	2 ± 1	6 ± 3	15 ± 11	4 ± 2	13 ± 8	$\mathbf{46 \pm 9}$	21 ± 9	36 ± 8
	task 2	2 ± 1	11 ± 4	1 ± 2	9 ± 3	16 ± 6	2 ± 3	$\mathbf{41 \pm 12}$	24 ± 1
	task 3	29 ± 3	57 ± 8	35 ± 23	67 ± 8	$\mathbf{75 \pm 6}$	31 ± 10	15 ± 4	57 ± 5
	task 4	6 ± 2	14 ± 5	7 ± 7	18 ± 6	27 ± 10	3 ± 2	$\mathbf{33 \pm 15}$	12 ± 6
	task 5	6 ± 3	16 ± 4	11 ± 8	18 ± 4	40 ± 8	$\mathbf{44 \pm 19}$	37 ± 20	$\mathbf{44 \pm 9}$
cube single play	overall	6 ± 2	13 ± 2	8 ± 4	13 ± 2	$\mathbf{30 \pm 8}$	2 ± 1	3 ± 2	21 ± 10
	task 1	5 ± 2	12 ± 4	6 ± 4	13 ± 2	$\mathbf{27 \pm 7}$	1 ± 1	3 ± 4	20 ± 7
	task 2	5 ± 2	13 ± 3	8 ± 6	13 ± 3	$\mathbf{30 \pm 10}$	3 ± 2	4 ± 4	17 ± 12
	task 3	7 ± 3	13 ± 4	14 ± 9	14 ± 4	$\mathbf{23 \pm 13}$	3 ± 3	4 ± 4	18 ± 9
	task 4	6 ± 2	15 ± 6	6 ± 4	11 ± 3	$\mathbf{37 \pm 19}$	2 ± 2	2 ± 2	21 ± 12
	task 5	5 ± 3	11 ± 3	5 ± 4	13 ± 4	$\mathbf{30 \pm 22}$	1 ± 2	1 ± 1	28 ± 20

The immediate contrast $C_{\pi_0} r$ is therefore retained exactly at every rank, including $k = 0$. We perform exactly one policy update, without actor iteration or feature refresh. The Boltzmann policies are

$$\pi_{k,\tau}(a \mid s) = \frac{\pi_0(a \mid s) \exp\{\tau \widehat{A}^{\pi_0}_{r,k}(s,a)\}}{\sum_b \pi_0(b \mid s) \exp\{\tau \widehat{A}^{\pi_0}_{r,k}(s,b)\}}, \qquad \tau \in \{4, 32, 256\}, \tag{26}$$

and the greedy policy selects $\arg\max_a \widehat{A}^{\pi_0}_{r,k}(s,a)$ with deterministic tie breaking. Grid World is evaluated at $k = 0, 5, \ldots, 45, 49$ and Four-room at $k = 0, 10, \ldots, 100, 103$.

All returns are computed exactly by solving the finite Bellman system from the same initial law μ:

$$J_r(\pi) = \mu^\top (I - \gamma P_\pi)^{-1} r_\pi. \tag{27}$$

The plotted quantity is the fraction of the available improvement achieved by the one-step policy,

$$\mathrm{AI}(\pi; r) = \frac{J_r(\pi) - J_r(\pi_0)}{J_r(\pi^*) - J_r(\pi_0)}, \tag{28}$$

where π^* and $J_r(\pi^*)$ are obtained by exact value iteration. Hence 0 denotes the reference-policy return and 1 denotes the optimal return and the score is not clipped.

Ablation: Learned representations used in the rank and Boltzmann temperature sweeps. The displayed learned-SUAF curves use the reward-free checkpoints selected by held-out operator error. Raw neural features are recalibrated on an independent positive-lag split using discounted feature-return calibration, separate left/right ridge penalties equal to 10^{-3} times the mean eigenvalue of the corresponding feature covariance, and a small-matrix SVD. The validation sweep selected zero additional singular-value shrinkage for both checkpoints. Checkpoint selection and calibration are reward-free. Downstream rewards are first used only in the linear coefficients $\langle \widehat{g}_i, r \rangle_{\widehat{\nu}}$ and the subsequent one-step update.

B.3 Spectral reward shaping in Taxi

Taxi as a sparse-reward stress test. We use the continuing Taxi environment as a deliberately difficult control problem. The state records the taxi position, passenger status, and destination, and the delivery reward is nonzero on only 4 of the 2400 reachable state-action pairs. Under the fixed behavior policy π_0, the positive delivery event has discounted occupancy mass only 1.48×10^{-5}.

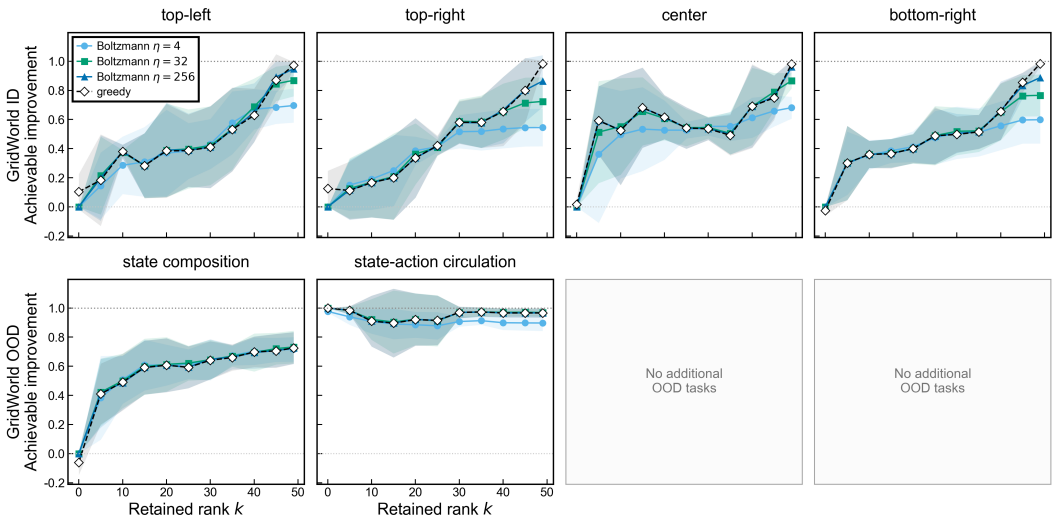


Figure 3: **One step policy update with exact SUAF in Grid World environment.** The results is calculated with 5 full-coverage random policies with 95% CI. This also works as an ablation of Boltzmann update with different temperature.

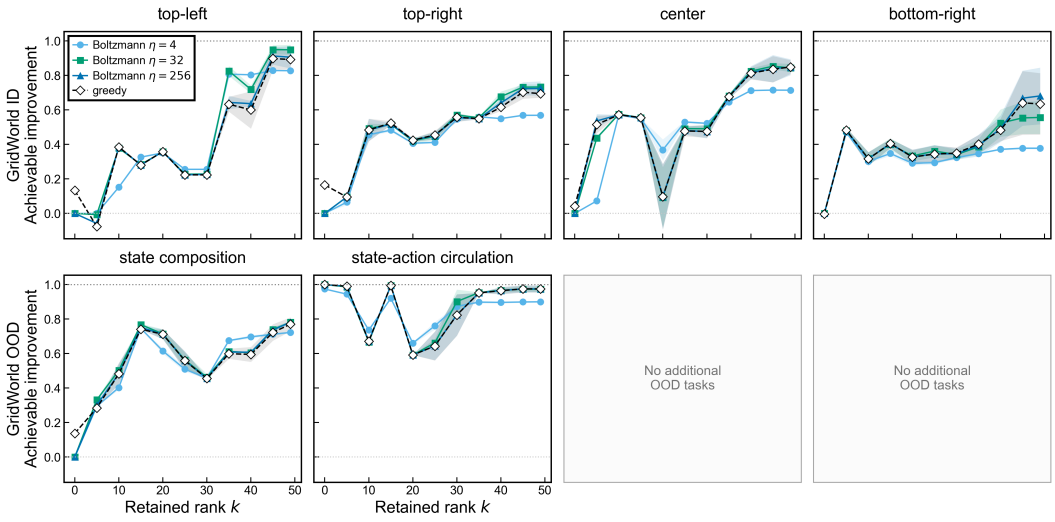


Figure 4: **One step policy update with learned SUAF in Grid World environment.** The results is calculated with 5 full-coverage random policies with 5 seeds learned SUAF with 95% CI. This also works as an ablation of Boltzmann update with different temperature.

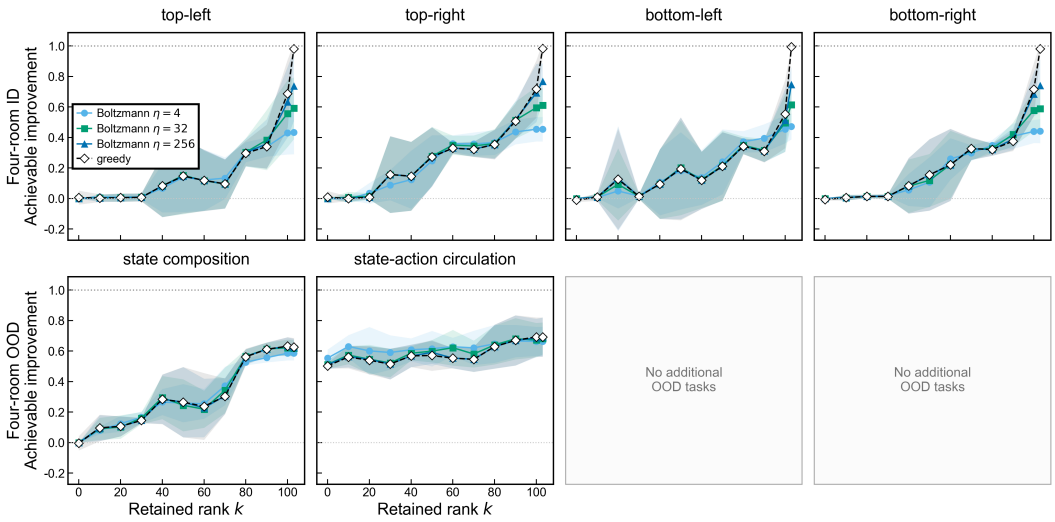


Figure 5: **One step policy update with exact SUAF in Four-room environment.** The results is calculated with 5 full-coverage random policies with 95% CI. This also works as an ablation of Boltzmann update with different temperature.

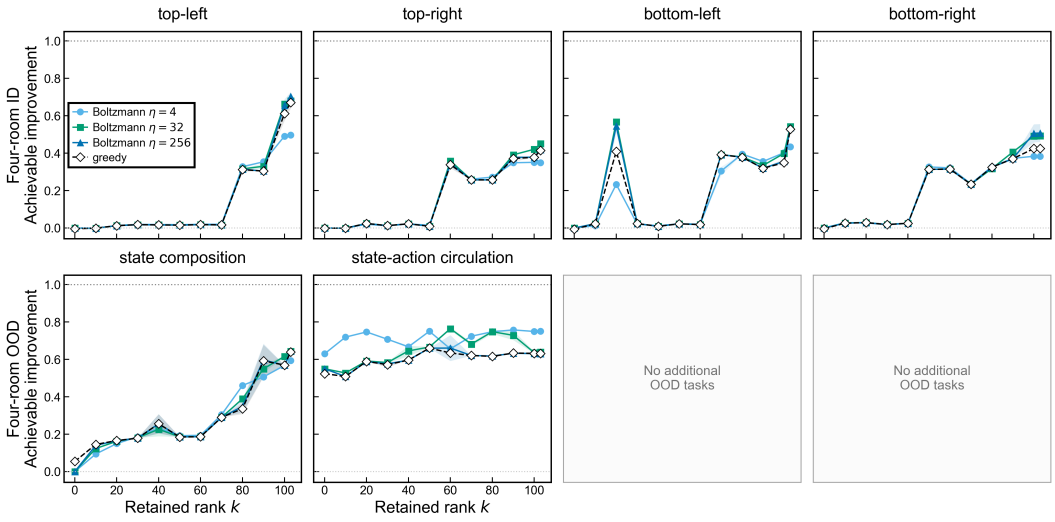


Figure 6: **One step policy update with learned SUAF in Four-room environment.** The results is calculated with 5 full-coverage random policies with 5 seeds learned SUAF with 95% CI. This also works as an ablation of Boltzmann update with different temperature.

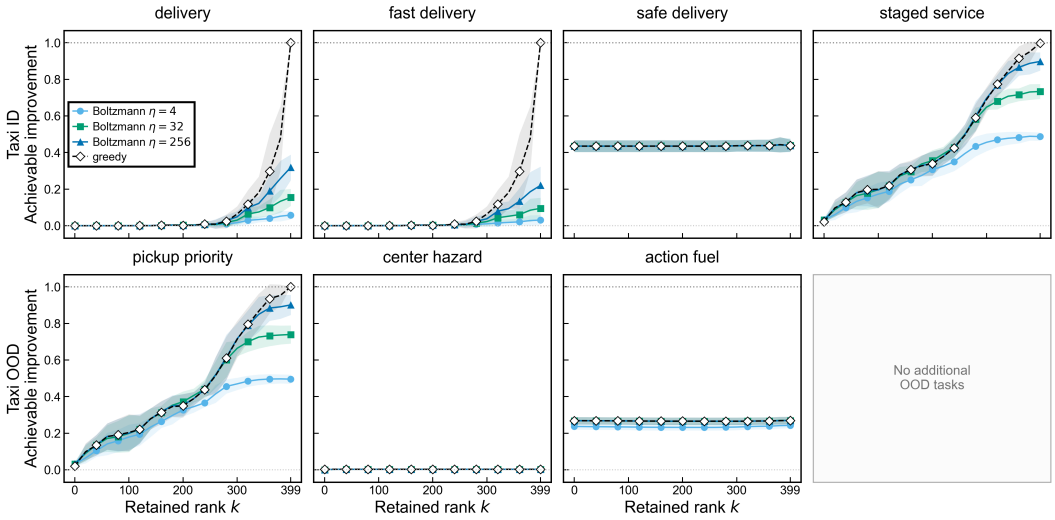


Figure 7: **One step policy update with exact SUAF in Taxi environment.** The results is calculated with 5 full-coverage random policies with 95% CI. This also works as an ablation of Boltzmann update with different temperature. The Taxi model outputs 499 coordinates as an overparameterization diagnostic, while the population DCDM has rank 399. Coordinates beyond rank 399 do not correspond to additional nonzero population SUAF modes.

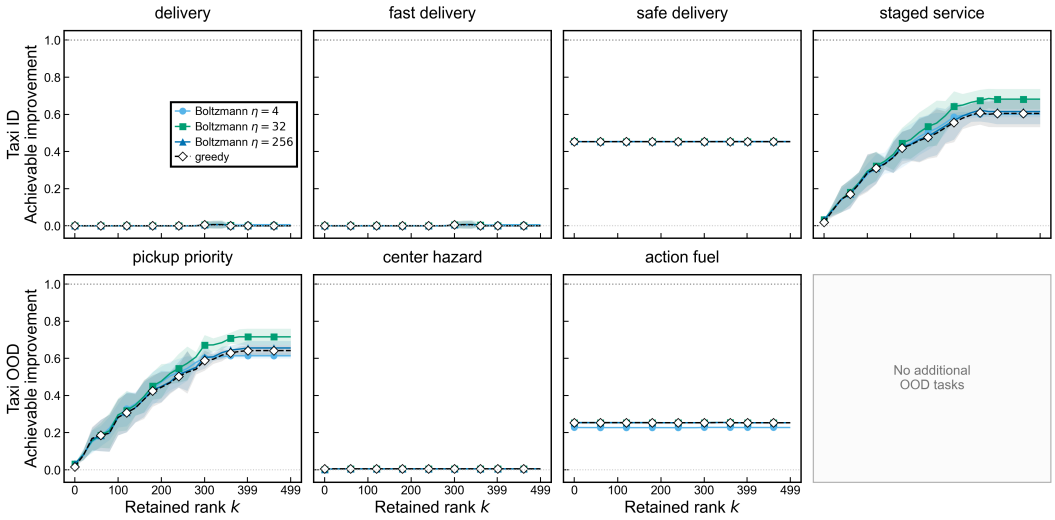


Figure 8: **One step policy update with learned SUAF in Taxi environment.** The results is calculated with 5 full-coverage random policies with 5 seeds with 95% CI. This also works as an ablation of Boltzmann update with different temperature.

Consequently, a reward-free representation must retain weak, long-horizon distinctions involving pickup, navigation, and drop-off in order to recover a useful delivery policy from a one-step readout.

This behavior is consistent with the discussion in Section A.5. Write the state reward averaged under the behavior policy as

$$r^{\pi_0}(s) := \sum_a \pi_0(a \mid s) r(s,a), \tag{29}$$

and let $\{(f_i, g_i, \sigma_i)\}_{i=1}^R$ denote the SUAF modes of π_0, where the right modes are orthonormal in $L^2(\nu^{+,\pi_0})$. Define

$$\beta_i(r) := \langle g_i, r^{\pi_0} \rangle_{\nu^{+,\pi_0}}. \tag{30}$$

For the unshaped delivery reward, only 0.88% of the delayed spectral energy $\sum_i \sigma_i^2 \beta_i(r)^2$ lies in the first 64 modes, whereas 66.0% lies after mode 256. Thus, low-rank or imperfectly learned representations can miss most of the signal even though the downstream reward itself has a simple semantic description.

An oracle spectral potential. To isolate the effect of spectral reward allocation, define the weighted projector onto the complete right singular subspace as $\Pi_G = GG^\top W_\nu$, where $W_\nu = \mathrm{diag}(\nu^{+,\pi_0})$, and let

$$r_\perp^{\pi_0} := (I - \Pi_G) r^{\pi_0}. \tag{31}$$

In continuing Taxi there are 400 reachable states and the population SUAF has rank $R = 399$. Because the right singular functions are centered and orthonormal under ν^{+,π_0}, they span the complete subspace of ν^{+,π_0}-centered state functions. The orthogonal complement is therefore only the constant direction, and

$$r_\perp^{\pi_0} = \mathbb{E}_{S \sim \nu^{+,\pi_0}} [r^{\pi_0}(S)] \, \mathbf{1}. \tag{32}$$

We can consequently define the head-m target reward as

$$r_{\text{head-}m}^{\pi_0}(s) := \mathbb{E}_{\nu^{+,\pi_0}} [r^{\pi_0}(S)] + \sum_{i=1}^m g_i(s) \beta_i(r). \tag{33}$$

Importantly, equation 33 requires neither the tail coefficients $\{\beta_i\}_{i>m}$ nor the tail singular functions. Given the reward and π_0, the state reward r^{π_0} is known and its weighted mean and the leading coefficients $\beta_i(r) = \langle g_i, r^{\pi_0} \rangle_{\nu^{+,\pi_0}}$ can be estimated directly from samples of ν^{+,π_0}. The constant orthogonal component is retained, while only the representable spectral tail is removed.

We then construct the state-only potential

$$\Phi_m = (I - \gamma P_{\pi_0})^{-1} (r^{\pi_0} - r_{\text{head-}m}^{\pi_0}) \tag{34}$$

and the shaped transition reward

$$r_{\Phi_m}(s,a,s') = r(s,a,s') + \gamma \Phi_m(s') - \Phi_m(s). \tag{35}$$

By construction,

$$r_{\Phi_m}^{\pi_0} = r_{\text{head-}m}^{\pi_0}, \qquad \left\langle g_i, r_{\Phi_m}^{\pi_0} \right\rangle_{\nu^{+,\pi_0}} = 0 \quad (i > m). \tag{36}$$

Potential shaping leaves the advantage and the set of optimal policies unchanged. Operationally, equation 34 does not move tail coefficients into head coefficients. Instead, it removes the tail from the delayed component and transfers its exact contribution to the immediate action-centered term.

Oracle-mode results. The population SUAF has rank 399. With the original sparse reward, oracle greedy one-step control remains near zero through rank 280 and reaches the optimal policy only at rank 399. The four learned full-rank SUAF models also obtain approximately zero improvement on the unshaped reward. In contrast, the oracle spectral potentials make the task accessible at the prescribed head rank for exact SUAF to recover the optimal policy, as is shown in Figure 9. Furthermore, using the learned representations, head-4, head-16, and head-64 shaping obtains normalized improvements 0.466 ± 0.047, 0.407 ± 0.021, and 0.396 ± 0.039, respectively (mean and 95% confidence interval), compared with approximately zero without shaping.

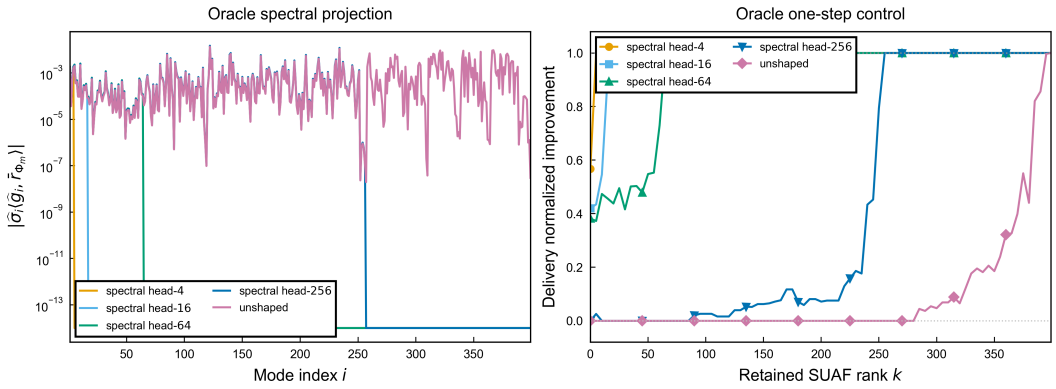


Figure 9: **Oracle spectral reward shaping in Taxi.** The left panel compares the spectral allocation of the sparse delivery reward and shaped variants for exact SUAF. The right panel reports one-step control on the original delivery objective. Curves or bars show the mean over four independently learned representations.

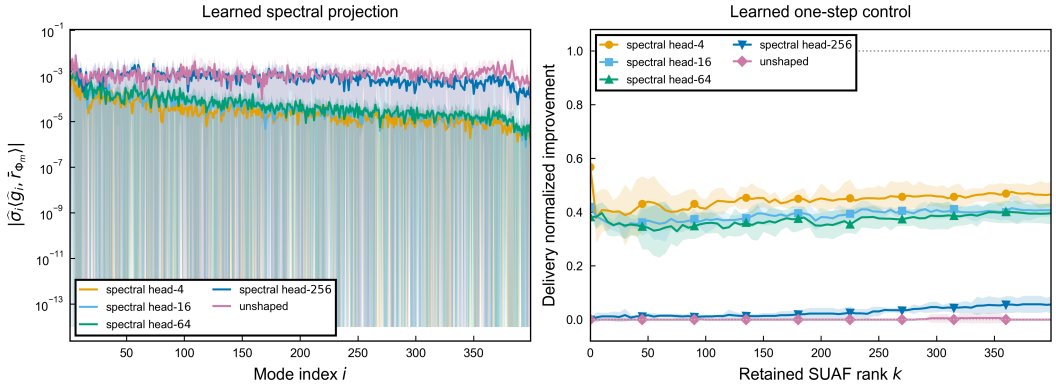


Figure 10: **Oracle spectral reward shaping in Taxi.** The left panel compares the spectral allocation of the sparse delivery reward and shaped variants for learned SUAF. The right panel reports one-step control on the original delivery objective. Curves or bars show the mean over four independently learned representations.

A learned spectral potential. We next replace the oracle right singular functions by the calibrated modes from each learned checkpoint and remove the constant component altogether:

$$\widehat{r}^{\pi_0}_{\text{head-}m}(s) = \sum_{i=1}^{m} \widehat{g}_i(s)\widehat{\beta}_i(r), \qquad \widehat{\beta}_i(r) = \langle \widehat{g}_i, r^{\pi_0} \rangle_{\nu^+,\pi_0} . \tag{37}$$

For this diagnostic we still use the known Taxi transition kernel to compute

$$\widehat{\Phi}_m = (I - \gamma P_{\pi_0})^{-1} \left(r^{\pi_0} - \widehat{r}^{\pi_0}_{\text{head-}m}\right). \tag{38}$$

Thus, the construction uses no oracle singular functions and no tail coefficients, but it continues to use oracle dynamics. At the matched truncation rank $k = m$, learned head-4, head-16, head-64, and head-256 shaping obtains normalized delivery improvements 0.246 ± 0.071, 0.169 ± 0.109, 0.107 ± 0.046, and 0.074 ± 0.056, respectively. The unshaped learned SUAF remains near zero. The smallest learned head therefore provides a clear but partial recovery, although it remains below the corresponding oracle-mode construction. Using all 399 learned modes again gives approximately zero improvement, showing that the gain comes from removing unreliable learned directions rather than from potential shaping alone. Correcting the learned Gram matrix before projection produces nearly identical results, so the remaining gap is not explained solely by imperfect orthogonality of the calibrated $\widehat{g}_i$.

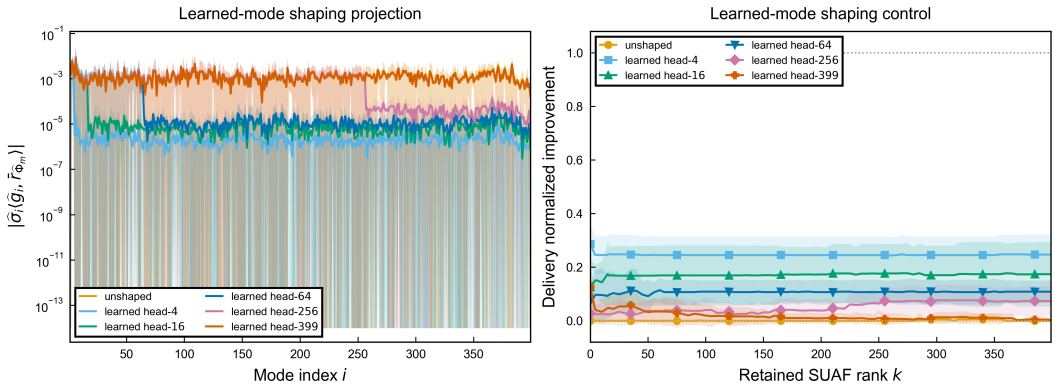


Figure 11: **Learned spectral reward shaping in Taxi.** The left panel compares the spectral allocation of the sparse delivery reward and shaped variants for learned SUAF. The right panel reports one-step control on the original delivery objective. Curves or bars show the mean over four independently learned representations.

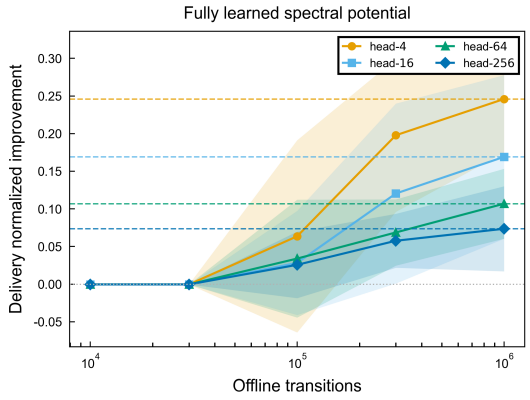


Figure 12: **Estimation of P_π with dataset.** Solid curves show means, and shaded regions indicate std across 5 independent runs.

A fully learned spectral potential. The preceding experiment still computes the potential with the population transition matrix. We remove this final oracle component using a fixed offline replay $\mathcal{D} = \{(S_n, A_n, S'_n)\}_{n=1}^N$. In the finite Taxi environment we first estimate

$$\widehat{P}_{\mathcal{D}}(s' \mid s, a) = \frac{\sum_{n=1}^N \mathbf{1}\{S_n = s, A_n = a, S'_n = s'\}}{\sum_{n=1}^N \mathbf{1}\{S_n = s, A_n = a\}},$$
$$\widehat{P}_{\pi_0,\mathcal{D}}(s, s') = \sum_a \pi_0(a \mid s)\widehat{P}_{\mathcal{D}}(s' \mid s, a). \tag{39}$$

No population transition probabilities are used in this estimator. Since $(I - \gamma\widehat{P}_{\pi_0,\mathcal{D}})^{-1} = I + \gamma\widehat{P}_{\pi_0,\mathcal{D}} + O(\gamma^2)$, we then compute the fully learned potential

$$\widetilde{\Phi}_{m,\mathcal{D}} = (I + \gamma\widehat{P}_{\pi_0,\mathcal{D}})\,(r^{\pi_0} - \widehat{r}^{\pi_0}_{\text{head-}m})\,. \tag{40}$$

This operation is equivalent to tabular fitted policy evaluation of the residual reward. In large or continuous state spaces, equation 40 can therefore be replaced by an offline FQE or TD critic without explicitly materializing a transition matrix.

As is shown in Figure 12, we evaluate replay sizes $N \in \{10^4, 3\times10^4, 10^5, 3\times10^5, 10^6\}$ and save the replay used for every seed. With 10^4 or 3×10^4 transitions, incomplete state-action coverage causes all variants to remain near zero. At 10^5 transitions, learned head-4 reaches 0.064 ± 0.128 normalized improvement, increasing to 0.198 ± 0.102 at 3×10^5. With 10^6 transitions it reaches

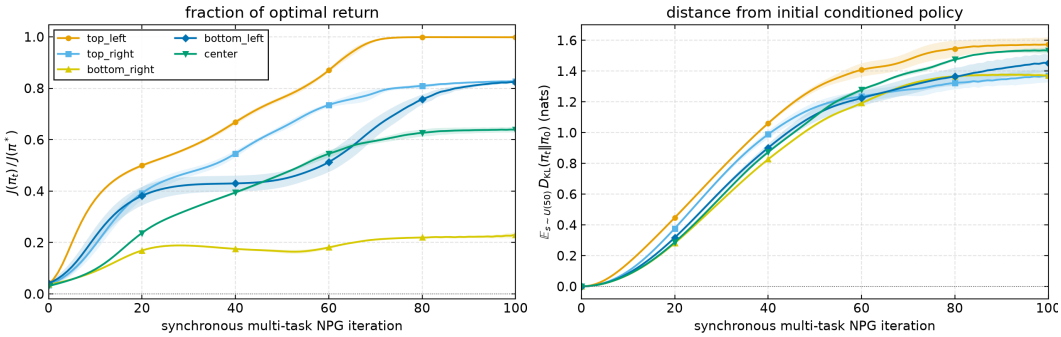


Figure 13: **Iterative multi-task policy improvement with SUAF in Grid World.** A shared reward-conditioned actor is optimized using synchronous multi-task natural policy gradient (NPG), with SUAF critics reused between refreshes triggered by the trust-region criterion. Solid curves show means, and shaded regions indicate std across 3 independent runs. **Left:** Return normalized by the task-specific optimal return, $J_r(\pi_t)/J_r(\pi_r^*)$. **Right:** KL divergence from the initial task-conditioned policy, averaged uniformly over the 50 states. Colors indicate reward tasks.

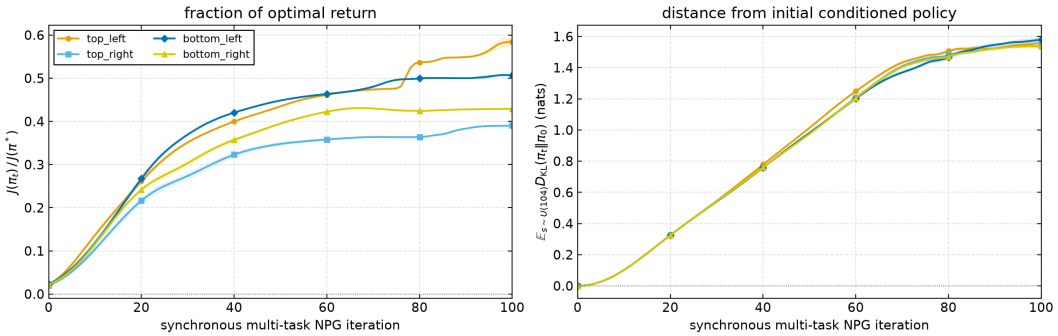


Figure 14: **Iterative multi-task policy improvement with SUAF in Four-rooms.** A shared reward-conditioned actor is optimized using synchronous multi-task natural policy gradient (NPG), with SUAF critics reused between refreshes triggered by the trust-region criterion. Solid curves show means, and shaded regions indicate std across 3 independent runs.**Left:** Return normalized by the task-specific optimal return, $J_r(\pi_t)/J_r(\pi_r^*)$. **Right:** KL divergence from the initial task-conditioned policy, averaged uniformly over the 104 states. Colors indicate reward tasks.

0.246 ± 0.071, numerically matching the learned-mode potential computed with the exact transition matrix. The same convergence occurs for head-16, head-64, and head-256, whose 10^6-transition results are respectively 0.169 ± 0.109, 0.107 ± 0.046, and 0.074 ± 0.056. At this sample size the dataset covers on average 2398.25 of the 2400 reachable state–action pairs, and the weighted potential error for head-4 decreases to 2.1×10^{-4}. Thus the benefit of learned spectral shaping survives after both the singular functions and the potential are learned from data.

B.4 Policy Update within Trust Region

After pretraining the SUAF features, we use the right features as a reward encoder. For a reward r, its task representation is

$$\beta(r) = \big(\langle g_1, r\rangle_{\nu^+,\pi_{\mathrm{ref}}}, \ldots, \langle g_k, r\rangle_{\nu^+,\pi_{\mathrm{ref}}}\big).$$

A single reward-conditioned actor is shared across all tasks:

$$\pi_\theta(a \mid s, r) = \pi_\theta(a \mid s, \beta(r)).$$

Thus, the actor receives both the current state and the SUAF reward representation, while all rewards are optimized jointly using the same parameter vector θ. This allows policy knowledge to be shared across tasks rather than training a separate actor for each reward.

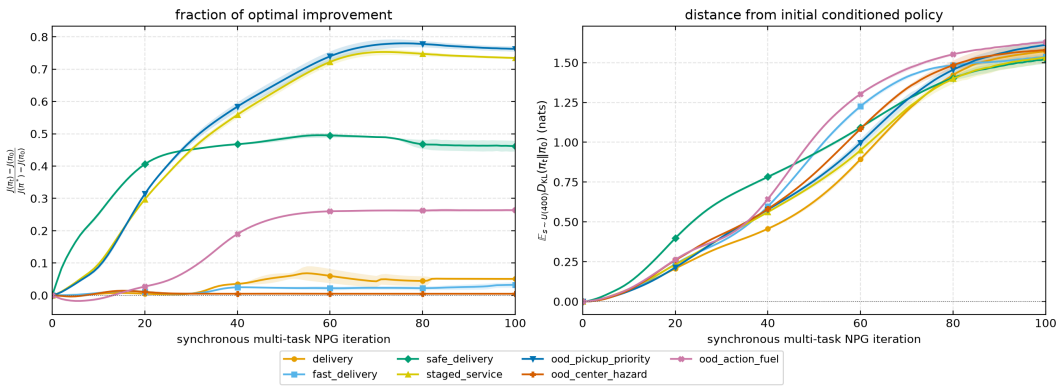


Figure 15: **Iterative multi-task policy improvement with SUAF in Taxi.** A shared reward-conditioned actor is optimized using synchronous multi-task natural policy gradient (NPG), with SUAF critics reused between refreshes triggered by the trust-region criterion. Solid curves show means, and shaded regions indicate std across 3 independent runs.**Left:** Return normalized by the task-specific optimal return, $J_r(\pi_t)/J_r(\pi_r^*)$. **Right:** KL divergence from the initial task-conditioned policy, averaged uniformly over the 104 states. Colors indicate reward tasks.

The SUAF critic is associated with a reference policy π_{ref}. During policy optimization, its features are reused while the current policy remains inside a trust region. We refresh the critic only when $D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) > \delta$. At a refresh, new on-policy transitions are collected, the critic features f and g are warm-started from their previous parameters and recalibrated, and the current policy becomes the new reference policy. The reward coordinate supplied to the actor may remain defined in the original pretrained coordinate system, providing a stable task representation while the policy-specific critic is updated.

We evaluate whether SUAF supports repeated policy improvement with a shared reward-conditioned actor and periodic critic refreshes. In Grid World, all five tasks achieve higher returns after optimization, although their learning trajectories differ. The top-left task reaches approximately the optimal return by iteration 80, while top-right and bottom-left reach approximately 83% of their respective optima by iteration 100. The center task reaches approximately 64%, whereas bottom-right improves more modestly to approximately 23% and exhibits a temporary decline during training. These results demonstrate that SUAF reward representations can support joint optimization of multiple tasks through a single actor, while also revealing substantial variation in the effectiveness of the updates across rewards.

In Four Rooms, all four tasks improve from their initial policies, reaching approximately 59%, 51%, 43%, and 39% of optimal return for top-left, bottom-left, bottom-right, and top-right, respectively. Most tasks improve rapidly during the early iterations and subsequently progress more slowly, with top-left showing additional gains late in training. The remaining optimality gaps indicate that repeated updates with a shared actor and refreshed SUAF critics do not uniformly recover optimal behavior within the evaluated budget.

The right panels show that the policies move progressively away from their initial task-conditioned policies, with the final mean KL divergence ranging from approximately 1.4 to 1.6 nats in Grid World and clustering around 1.5–1.6 nats in Four Rooms. Similar cumulative policy displacements can produce substantially different returns, emphasizing that update magnitude alone does not determine improvement. These plots measure divergence from the initial policy under a uniform state distribution. Local reuse is instead governed by the refresh criterion relative to the current reference policy. Together, the results demonstrate the feasibility of combining SUAF-based task conditioning with iterative policy optimization and critic refreshes, while exposing task-dependent limitations in the resulting performance.

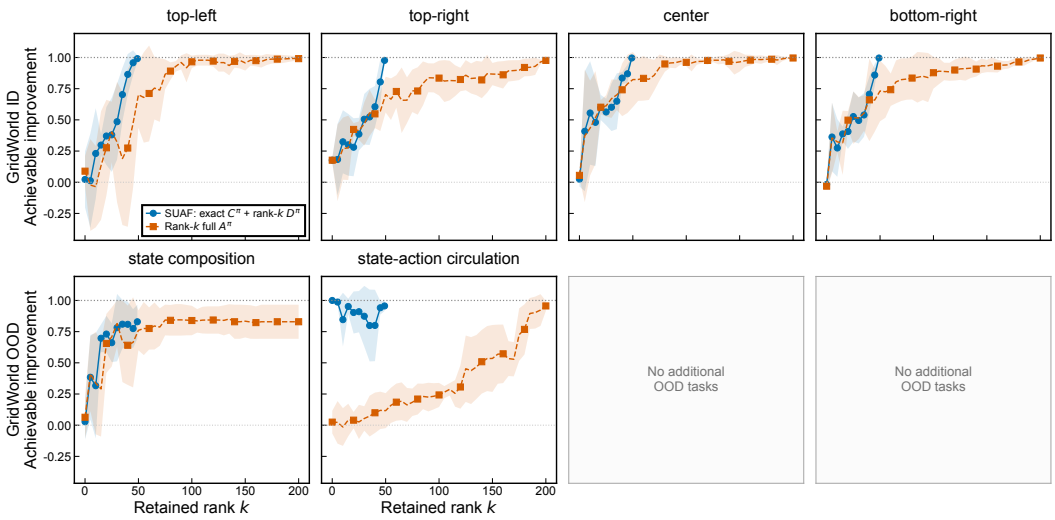


Figure 16: **Comprasion between SUAF and SVD in one-step improvement in GridWorld.** The results are reported as mean with std of 5 independent runs.

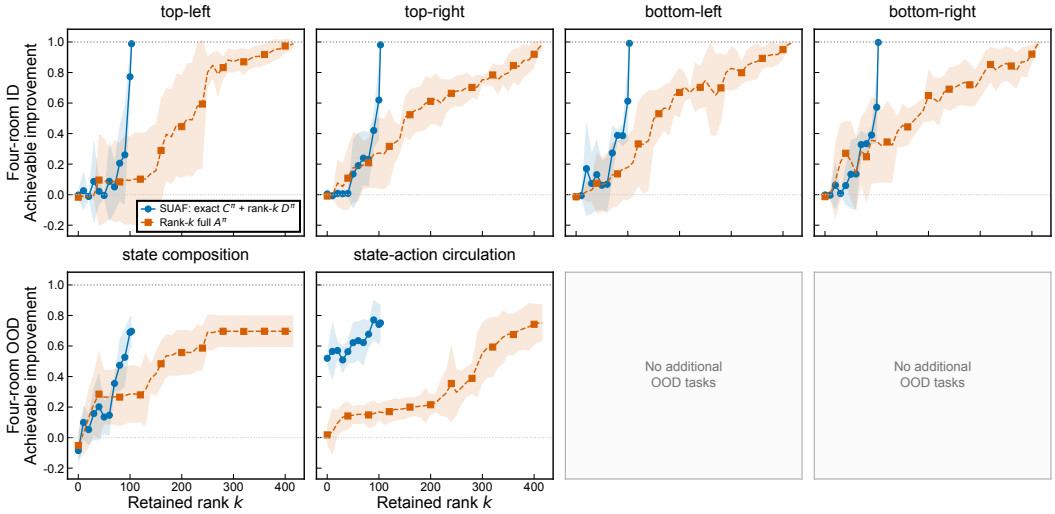


Figure 17: **Comprasion between SUAF and SVD in one-step improvement in Four-room.** The results are reported as mean with std of 5 independent runs.

C Ablation Studies

C.1 Does Separating the Immediate Contrast Matter for Control?

Theorem 3.3 shows that directly compressing the full advantage operator must spend representation capacity on the immediate action contrast, whereas SUAF retains this known term exactly and uses its spectral modes only for delayed credit. We next examine whether this distinction matters for downstream control, rather than only for operator reconstruction error.

Compared operators. For a fixed reference policy π_0, write the advantage operator as

$$\mathcal{A}^{\pi_0} = C_{\pi_0} + \mathcal{D}^{\pi_0}, \qquad \mathcal{D}^{\pi_0} = \frac{\gamma}{1-\gamma}\mathcal{T}^{+,\pi_0}. \tag{41}$$

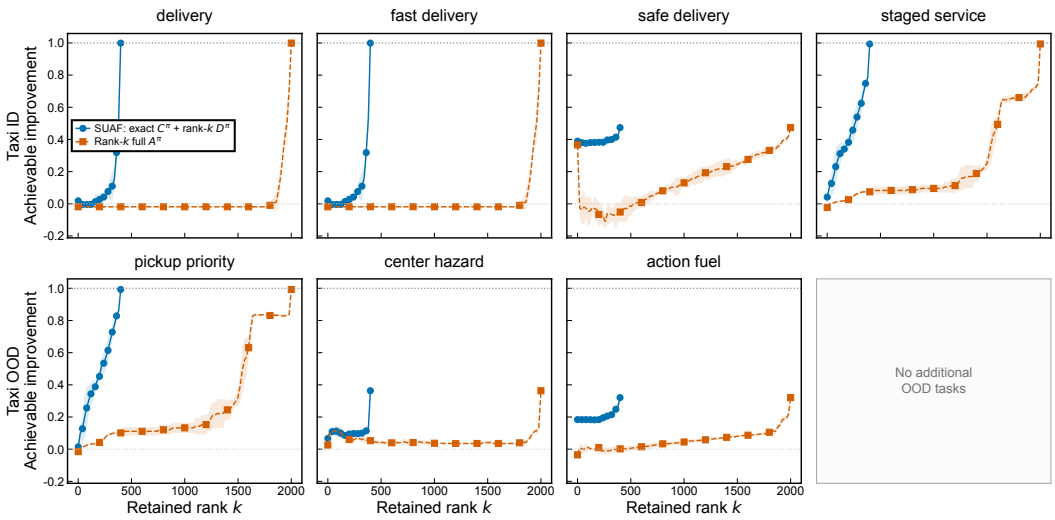


Figure 18: **Comprasion between SUAF and SVD in one-step improvement in Taxi.** The results are reported as mean with std of 5 independent runs.

We compare two population, reward-independent rank-k constructions. The SUAF approximation retains the immediate contrast exactly and truncates only the delayed operator:

$$\widehat{A}^{\mathrm{SUAF}}_{r,k} = C_{\pi_0} r + \mathcal{D}^{\pi_0}_k r, \tag{42}$$

where $\mathcal{D}^{\pi_0}_k$ is the leading rank-k weighted SVD approximation of $\mathcal{D}^{\pi_0}$. The ablation instead directly truncates the full advantage operator:

$$\widehat{A}^{\mathrm{Full}}_{r,k} = \left(\mathcal{A}^{\pi_0}\right)_k r, \tag{43}$$

where $\left(\mathcal{A}^{\pi_0}\right)_k$ is the leading rank-k approximation of the full operator under the same weighted $L^2(d^{\pi_0})$–$L^2(\nu^{+,\pi_0})$ geometry.

Here, k denotes the number of retained spectral modes that must be represented. The exact term $C_{\pi_0} r$ in SUAF introduces no learned parameters: it is computed directly from the downstream reward and the known reference policy. Thus, the comparison allocates the same *learned spectral capacity* to both methods.

Control protocol. To isolate the structural effect from representation-learning error, we compute both decompositions from the exact finite MDP. For every rank k and downstream reward r, we construct a one-step greedy policy

$$\pi_k(a \mid s) = \mathbf{1}\left\{ a \in \arg\max_{a'} \widehat{A}^{\pi_0}_{r,k}(s, a') \right\}, \tag{44}$$

with fixed tie breaking. We measure its achievable improvement by

$$\mathrm{AI}(\pi_k) = \frac{J_r(\pi_k) - J_r(\pi_0)}{J_r(\pi^\star_r) - J_r(\pi_0)}. \tag{45}$$

An achievable improvement of one therefore indicates that the one-step policy reaches the optimal return, while zero corresponds to no improvement over the reference policy.

We repeat the population experiment over five independently sampled full-support reference policies and report the mean and Student-t 95% confidence interval. The uncertainty therefore measures sensitivity to the reference policy, rather than stochastic error in the exact SVD.

Results. Figure 16, 17, 18 show that the operator-level spectral floor has a direct consequence for control. The delayed operators have ranks 49, 103, and 399 in GridWorld, Four-room, and Taxi, respectively. At these same rank budgets, SUAF has already recovered the exact advantage of the

Table 6: Full-rank greedy one-step control with frozen FB representations. The exact-C_π variants (labeled delayed) use $C_\pi r + \widehat{D}^\pi r$ and original ones use $\widehat{A}^\pi r$. Values are normalized improvement from the reference policy. Result is reported with 5 independent runs.

Environment	task	FB	FB + delayed	One-Step FB	One-Step FB + delayed	Ours w/o delayed	Ours
GridWorld	top-left	$0.979_{\pm 0.038}$	$0.978_{\pm 0.032}$	$0.767_{\pm 0.126}$	$0.759_{\pm 0.115}$	$0.746_{\pm 0.079}$	$0.893_{\pm 0.057}$
	state composition	$0.890_{\pm 0.106}$	$0.801_{\pm 0.104}$	$0.777_{\pm 0.044}$	$0.772_{\pm 0.021}$	$0.640_{\pm 0.040}$	$0.771_{\pm 0.017}$
	state-action circulation	$0.058_{\pm 0.068}$	$0.554_{\pm 0.045}$ (↑)	$0.072_{\pm 0.041}$	$0.685_{\pm 0.054}$ (↑)	$0.860_{\pm 0.082}$	$0.975_{\pm 0.021}$
Four-room	top-left	$0.508_{\pm 0.006}$	$0.510_{\pm 0.008}$	$0.022_{\pm 0.003}$	$0.020_{\pm 0.004}$	$0.561_{\pm 0.103}$	$0.671_{\pm 0.031}$
	state composition	$0.660_{\pm 0.117}$	$0.595_{\pm 0.094}$	$0.321_{\pm 0.026}$	$0.322_{\pm 0.021}$	$0.601_{\pm 0.036}$	$0.640_{\pm 0.017}$
	state-action circulation	$0.111_{\pm 0.068}$	$0.398_{\pm 0.033}$ (↑)	$0.038_{\pm 0.017}$	$0.483_{\pm 0.024}$ (↑)	$0.582_{\pm 0.026}$	$0.632_{\pm 0.007}$
walker	flip	$277_{\pm 100}$	$262_{\pm 90}$	$388_{\pm 30}$	$360_{\pm 20}$	$147_{\pm 120}$	$357_{\pm 42}$
	run	$194_{\pm 20}$	$262_{\pm 90}$	$198_{\pm 30}$	$200_{\pm 18}$	$187_{\pm 102}$	$210_{\pm 20}$

reference policy, whereas the rank-k decomposition of the full advantage operator still omits many immediate action-contrast directions.

Overall, these results complement Theorem 3.3: the immediate-term spectral floor is not merely an operator-norm artifact, but can substantially reduce downstream policy improvement under a fixed representation budget.

C.2 Does other methods benefit from delayed advantage?

To test the source of performance improvement, we adapt FB and One-Step FB to the same delayed-operator target as SUAF, rather than training a full successor representation and subtracting an estimated identity component afterwards. Let $X_t = (S_t, A_t)$ under a fixed reference policy π_0, and sample a strictly positive geometric lag,

$$L \sim \mathrm{Geom}(1-\gamma), \qquad L \in \{1, 2, \ldots\}, \qquad Y^+ = X_{t+L}. \tag{46}$$

Thus the positive pairs used by the FB-style representation losses are drawn from

$$M^{+,\pi_0} = (1-\gamma) P_{\pi_0} (I - \gamma P_{\pi_0})^{-1}, \tag{47}$$

instead of from the usual full geometric successor kernel, which includes the lag-zero atom. We otherwise retain the original FB or One-Step FB network, optimizer, batch size, and orthonormalization/LSIF objective. In particular, the diagonal positive-pair term is evaluated on (X, Y^+), while the factual one-step transition remains available for the bootstrap term. Consequently, the resulting reward-conditioned score is trained to represent the delayed action-dependent component, denoted by $\widehat{D}^{\pi_0}_{\mathrm{base}} r$. At evaluation, we do not estimate the immediate term. We compute it exactly from the known reference policy,

$$(C_{\pi_0} r)(s,a) = r(s,a) - \sum_{\tilde{a}} \pi_0(\tilde{a} \mid s) r(s, \tilde{a}), \tag{48}$$

and use the critic

$$\widehat{A}^{\pi_0}_{\mathrm{base}} r = C_{\pi_0} r + \widehat{D}^{\pi_0}_{\mathrm{base}} r. \tag{49}$$

The equation 48 requires neither a transition model nor reward-specific representation training. It is exact for every test reward once π_0 is fixed. This intervention is non-vacuous only for an action-dependent reward: if $r(s,a) = r(s)$, then $C_{\pi_0} r = 0$ and the delayed critic and equation 49 coincide identically.

Why the ablation is restricted to FB and One-Step FB. This experiment requires a baseline whose representation-learning objective can be modified in a controlled way by replacing its successor positive examples with the pairs in equation 46. FB and One-Step FB have explicit forward–backward successor objectives with such a positive-pair/bootstrapping decomposition, so the change preserves their architecture and learning rule while changing only the operator being learned.

The remaining baselines do not admit an analogous one-line intervention. Laplacian and BYOL-γ learn state geometry rather than a specified successor operator; ICVF and HILP use task- or intention-conditioned value objectives; and TD-JEPA uses a representation-specific prediction objective. None exposes an explicit successor positive-pair term whose lag-zero mass can be removed while retaining the original objective. Adding a new reward-conditioned delayed TD or successor-feature head to each method would introduce an additional model, loss, and optimization budget.

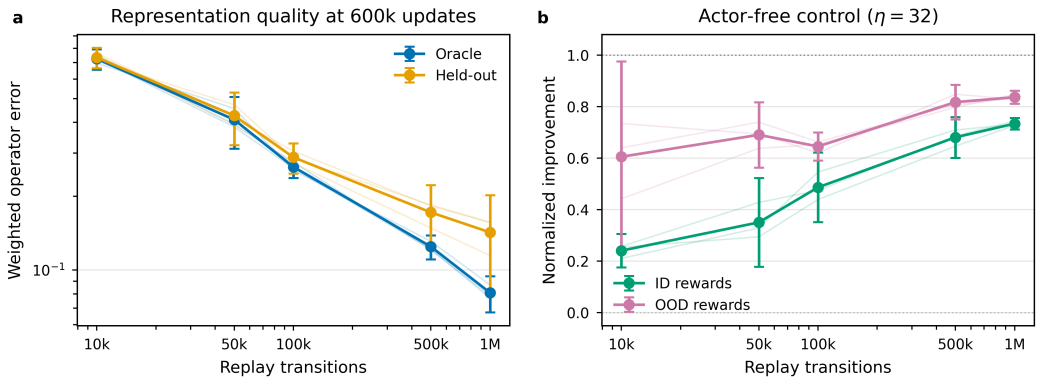


Figure 19: **Effects of training dataset on feature learning in GridWorld.** Results are reported with mean ± 95% CI with 3 independent runs.

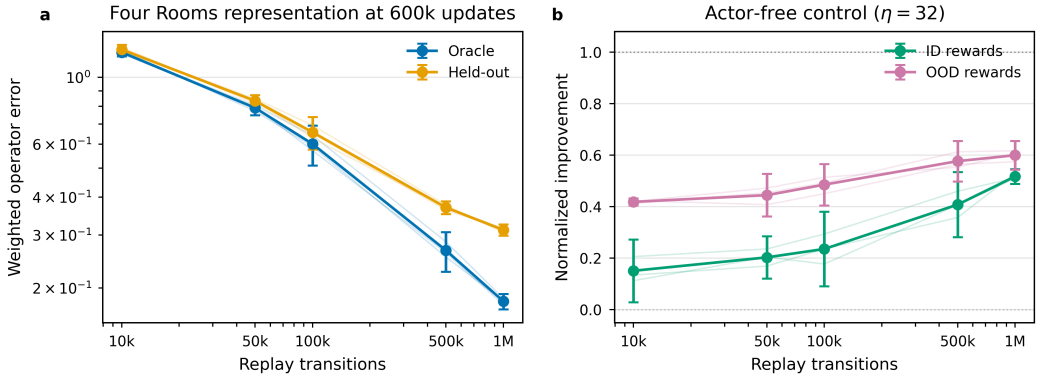


Figure 20: **Effects of training dataset on feature learning in Four-room.** Results are reported with mean ± 95% CI with 3 independent runs.

Such a comparison would measure that added head rather than the original baseline. We therefore apply the exact-C_{π_0} delayed-operator ablation only to FB and One-Step FB.

Delayed ablation. Table 6 isolates the effect of separating the immediate reward contrast $C_\pi r$ from the delayed component. For state-only rewards, $C_\pi r = 0$, so adding an exact C_π term leaves performance unchanged, as expected. For state-action rewards, however, explicitly computing $C_\pi r$ yields large gains for the baselines, confirming that direct low-rank compression of the full advantage operator suffers from the immediate-term spectral floor. Crucially, even when the baselines are given the same exact $C_\pi r$, SUAF still shows great advantage over them, showing that its learned delayed spectral representation is itself stronger. Overall, the ablation supports the core design: exact immediate contrast plus a low-rank spectral model of the delayed operator.

How does the dataset size affect performance?

C.3 How does estimation of behavior policy affect the final results?

We use the cross-episode held-out action R^2 to measure the quality of the kNN estimate of the behavior policy $\pi_b(a \mid s)$, where higher values indicate that the neighboring actions more accurately predict the behavior action at the current state. The R^2 is computed on held-out data from different episodes as

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \|a_i - \widehat{a}_k(s_i)\|_2^2}{\sum_{i=1}^{N} \|a_i - \bar{a}_{\text{ref}}\|_2^2},$$

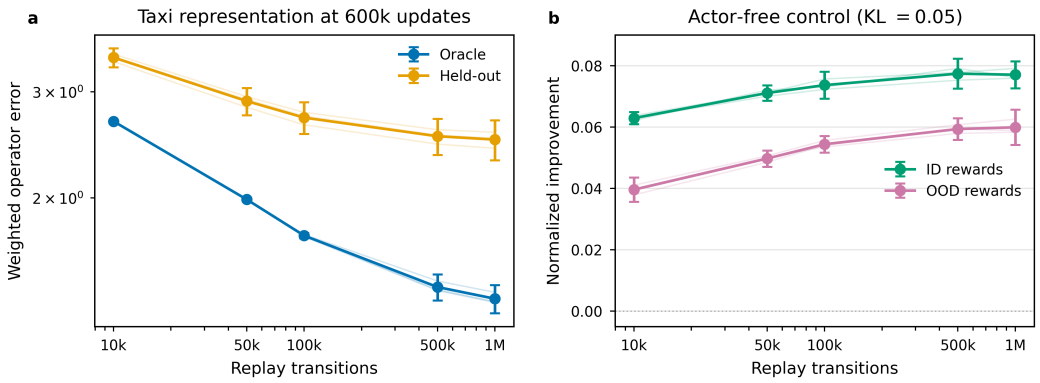


Figure 21: **Effects of training dataset on feature learning in Taxi.** Results are reported with mean $\pm$ 95% CI with 3 independent runs.

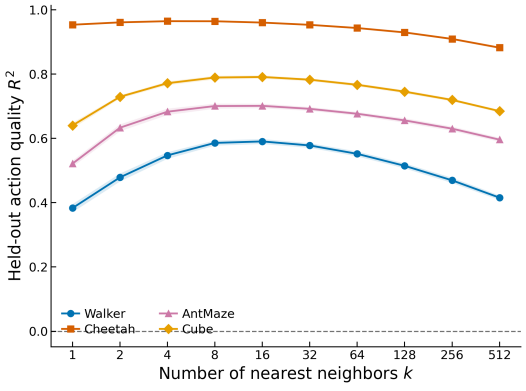


Figure 22: **Quality of kNN estimation measured by held-out action** R^2. The results are reported by 95% CI of 5 independent episode split.

Table 7: **Ablations on kNN action numbers k.** The full zero-shot evaluation results on 8 ExORL tasks and 10 OGBench tasks. The results are reported as *mean* over 3 seeds with *std* while $k = 16$ using the original results.

Domain	Task	$k = 1$	$k = 4$	$k = 16$	$k = 64$	$k = 256$
ExORL: Locomotion						
walker	overall	117 ± 12	269 ± 24	$\mathbf{445 \pm 39}$	266 ± 24	58 ± 11
	flip	43 ± 6	252 ± 12	$\mathbf{357 \pm 42}$	281 ± 2	38 ± 7
	run	73 ± 11	184 ± 13	$\mathbf{210 \pm 20}$	108 ± 33	85 ± 23
	stand	172 ± 20	256 ± 46	$\mathbf{670 \pm 100}$	286 ± 12	63 ± 23
	walk	181 ± 7	383 ± 3	$\mathbf{542 \pm 41}$	391 ± 56	46 ± 20
cheetah	overall	395 ± 21	376 ± 28	$\mathbf{375 \pm 21}$	278 ± 21	152 ± 13
	run	107 ± 21	94 ± 33	$\mathbf{100} \pm 12$	76 ± 25	28 ± 17
	run backward	149 ± 16	161 ± 12	$\mathbf{159 \pm 70}$	96 ± 8	57 ± 11
	walk	497 ± 53	447 ± 44	$\mathbf{420 \pm 32}$	199 ± 30	58 ± 14
	walk backward	826 ± 75	804 ± 80	$\mathbf{820 \pm 232}$	742 ± 26	256 ± 37
OGBench: Navigation and manipulation						
antmaze large	overall	2 ± 1	17 ± 5	$\mathbf{35 \pm 5}$	17 ± 6	5 ± 3
	task 1	0 ± 0	10 ± 3	$\mathbf{36 \pm 8}$	20 ± 8	0 ± 0
	task 2	0 ± 0	10 ± 4	$\mathbf{24 \pm 1}$	11 ± 3	6 ± 4
	task 3	4 ± 1	27 ± 6	$\mathbf{57 \pm 5}$	25 ± 14	6 ± 2
	task 4	0 ± 0	5 ± 2	$\mathbf{12 \pm 6}$	3 ± 1	0 ± 0
	task 5	7 ± 2	33 ± 19	$\mathbf{44 \pm 9}$	28 ± 10	11 ± 8
cube single play	overall	2 ± 1	14 ± 6	$\mathbf{21 \pm 10}$	13 ± 5	3 ± 2
	task 1	2 ± 1	14 ± 9	$\mathbf{20 \pm 7}$	16 ± 8	4 ± 3
	task 2	2 ± 1	12 ± 4	$\mathbf{17 \pm 12}$	10 ± 5	4 ± 2
	task 3	3 ± 2	13 ± 8	$\mathbf{18 \pm 9}$	9 ± 5	1 ± 1
	task 4	0 ± 0	15 ± 4	$\mathbf{21 \pm 12}$	12 ± 10	1 ± 0
	task 5	4 ± 2	17 ± 9	$\mathbf{28 \pm 20}$	18 ± 9	3 ± 1

where

$$\widehat{a}_k(s_i) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(s_i)} a_j$$

is the average of the k nearest-neighbor actions for state s_i, and $\bar{a}_{\text{ref}}$ is the average action over the entire reference set. Thus:

- $R^2 = 1$: perfect prediction of the held-out action;
- $R^2 = 0$: equivalent to always predicting the global mean action;
- $R^2 < 0$: worse than the global mean action.

The reference and held-out queries come from different episodes, which avoids leakage from adjacent trajectory samples and prevents inflated results.

All four environments exhibit a clear bias–variance trade-off: too few neighbors lead to high estimation variance, whereas too many neighbors mix in samples whose states differ substantially. Walker, AntMaze, and Cube perform best around $k = 16$, while Cheetah achieves its best performance at $k = 4 \sim 8$. Overall, $k = 16$ is a reasonably robust unified choice across environments. Furthermore, we train with different number of kNN actions and found the result agree with the outcome.

D THEORY PROOFS

Table 8 pairs each main-text theorem with its proof in this appendix. The assumptions used throughout are collected in Appendix D.1.

Table 8: Main-text theorems and their corresponding proofs.

Theorem	Result	Proof
Theorem 3.1	Universal advantage function representation	Appendix D.4
Theorem 3.2	Average-case optimality of the rank-k representations	Appendix D.3
Theorem 3.3	Removing the immediate-term spectral floor	Appendix D.2
Theorem 4.1	Boltzmann policy improvement with truncated SUAF	Appendix D.6
Theorem 4.2	Policy improvement using transferred SUAF	Appendix D.9
Theorem 4.3	Policy-gradient error with truncated SUAF	Appendix D.10

D.1 ASSUMPTIONS

In this work, unless stated clearly, we consider a fixed discounted MDP with a known discount factor $0 < \gamma < 1$ and a reference policy π. Our construction uses the following assumptions:

1. **Access to the reference policy.** We can evaluate $\pi(a \mid s)$ and sample actions from $\pi(\cdot \mid s)$. This allows us to center functions across actions at a given state and to define the state–action transitions under π.
2. **Access to transition data.** We have trajectories under π, or appropriately weighted transition data, from which the discounted occupancy d^π, one-step pairs, and positive-lag pairs can be estimated. The transition probabilities need not be known explicitly.
3. **Access to downstream rewards.** Representation learning does not use task rewards. When adapting to a task, its bounded reward function $r(s, a)$ can be evaluated at the state-action pairs needed to compute the immediate reward contrast and the reward-feature coefficients.
4. **Well-defined feature distributions.** The weighted feature construction is restricted to the supports of d^π and $\nu^{+,\pi}$, where the required expectations and second moments are finite.

The exact finite-MDP policy-recovery result additionally assumes the full-support and exact-representation conditions stated in Theorem D.5.

D.2 Reason of exclude the immediate term and Proof of Theorem 3.3

The immediate reward contrast has a known high-rank structure that need not be learned from future transitions. The following result makes this precise in a common input-output geometry.

Lemma D.1 (Spectral floor induced by the immediate term)**.** *Throughout this lemma, both rewards and advantages belong to $\mathcal{H} = L^2(d^\pi)$, and all adjoints, singular values, and operator norms refer to this space. Consider a finite MDP with $0 < \gamma < 1$ and $d^\pi(s,a) > 0$ for all state-action pairs. Let P_π be the state-action transition operator and define the unnormalized-return advantage operator and its delayed component by*

$$\mathcal{A}_\pi = C_\pi(I - \gamma P_\pi)^{-1}, \qquad \mathcal{D}_\pi = \mathcal{A}_\pi - C_\pi = \frac{\gamma}{1-\gamma}\mathcal{T}^{+,\pi}.$$

Set $q = \operatorname{rank}(C_\pi) = |\mathcal{S}|(|\mathcal{A}| - 1)$ and let $\tau_1 \geq \cdots \geq \tau_q \geq 0$ be the singular values of $\mathcal{D}_\pi$, padded with zeros. Then $\mathcal{A}_\pi$ has exactly q nonzero singular values, given by

$$\sigma_i(\mathcal{A}_\pi) = \sqrt{1 + \tau_i^2}, \qquad i = 1, \ldots, q. \tag{50}$$

Consequently, for $0 \leq k < q$,

$$\inf_{\operatorname{rank}(K) \leq k} \|\mathcal{A}_\pi - K\|_{\mathrm{op}}^2 = 1 + \tau_{k+1}^2 \geq 1, \tag{51}$$

$$\inf_{\operatorname{rank}(K) \leq k} \|\mathcal{A}_\pi - (C_\pi + K)\|_{\mathrm{op}}^2 = \tau_{k+1}^2. \tag{52}$$

Proof. Since $d^\pi(s,a) = d_S^\pi(s)\pi(a \mid s)$, the action-centering operator C_π is the orthogonal projection onto $\mathcal{H}_C = \{u : \mathbb{E}_{a\sim\pi(\cdot|s)}u(s,a) = 0 \text{ for all } s\}$. The next action is sampled from π, so $P_\pi C_\pi = 0$. Hence $(I - \gamma P_\pi)^{-1} C_\pi = C_\pi$ and

$$\mathcal{D}_\pi C_\pi = 0, \qquad C_\pi \mathcal{D}_\pi = \mathcal{D}_\pi.$$

Using $C_\pi^* = C_\pi = C_\pi^2$ therefore gives

$$\mathcal{A}_\pi \mathcal{A}_\pi^* = (C_\pi + \mathcal{D}_\pi)(C_\pi + \mathcal{D}_\pi)^* = C_\pi + \mathcal{D}_\pi \mathcal{D}_\pi^*.$$

On $\mathcal{H}_C$, this is $I + \mathcal{D}_\pi \mathcal{D}_\pi^*$. On $\mathcal{H}_C^\perp$, it is zero. This proves equation 50. The two approximation identities follow from best rank-k approximation in operator norm, applying it to $\mathcal{A}_\pi$ and $\mathcal{D}_\pi$, respectively. □

Therefore, even when the delayed component is low rank, the full advantage operator remains a spectral floor of one on the q action-centered directions. Retaining $C_\pi r$ exactly avoids spending learned representation capacity on this known immediate structure. The comparison allocates rank k to the learned component. $C_\pi + K$ itself need not have rank k. The unit floor concerns unnormalized returns in the common $L^2(d^\pi)$ geometry: the immediate term in the normalized successor measure instead has coefficient $1 - \gamma$. Accordingly, the result motivates separating the immediate term without asserting a universal spectral-decay ordering between matrices normalized using different marginals. Note that Ollivier (2025) uses the spectrum analysis of $\mathcal{A}_\pi$, which introduce a non-zero error in the approximation of $\mathcal{A}_\pi$.

Based on this, we here prove the Theorem 3.3.

Theorem D.2 (Removing the immediate-term spectral floor)**.** *Consider a finite MDP with $0 < \gamma < 1$ and strictly positive d^π and $\nu^{+,\pi}$. Let*

$$\alpha_\pi = \min_s \frac{d_S^\pi(s)}{\nu_S^{+,\pi}(s)}, \qquad q = |\mathcal{S}|(|\mathcal{A}| - 1).$$

For every $0 \leq k < q$,

$$\inf_{\operatorname{rank}(K) \leq k} \sup_{\|r\|_{\nu^{+,\pi}} \leq 1} \|A_r^\pi - Kr\|_{d^\pi}^2 \geq \alpha_\pi + \left(\frac{\gamma}{1-\gamma}\right)^2 \left(\sigma_{k+1}^{+,\pi}\right)^2, \tag{53}$$

where singular values beyond the rank of $\mathcal{T}^{+,\pi}$ are defined to be zero. In contrast, retaining $C_\pi r$ exactly and truncating only the future term incurs worst-case error without α_π as shown in Theorem 3.2. If $\nu^{+,\pi} = d^\pi$, then $\alpha_\pi = 1$ and equation 6 holds with equality.

Proof. All adjoints and singular values below are taken between $L^2(\nu^{+,\pi})$ and $L^2(d^\pi)$. Write

$$\mathcal{A}_\pi = C_\pi + \mathcal{D}_\pi, \qquad \mathcal{D}_\pi = \frac{\gamma}{1-\gamma}\mathcal{T}^{+,\pi}.$$

Both d^π and $\nu^{+,\pi}$ have conditional action distribution π. Consequently, the adjoint of $C_\pi : L^2(\nu^{+,\pi}) \to L^2(d^\pi)$ is

$$C_\pi^* u(s,a) = \frac{d_S^\pi(s)}{\nu_S^{+,\pi}(s)}(C_\pi u)(s,a).$$

Its range consists of functions with zero policy-conditional action mean. Since P_π annihilates every such function,

$$\mathcal{D}_\pi C_\pi^* = 0, \qquad C_\pi \mathcal{D}_\pi^* = 0.$$

Hence

$$\mathcal{A}_\pi \mathcal{A}_\pi^* = C_\pi C_\pi^* + \mathcal{D}_\pi \mathcal{D}_\pi^*.$$

Let $\mathcal{H}_C = \mathrm{range}(C_\pi)$, viewed as a subspace of $L^2(d^\pi)$. This space has dimension q, and on $\mathcal{H}_C$,

$$C_\pi C_\pi^* = W_\pi, \qquad (W_\pi u)(s,a) = \frac{d_S^\pi(s)}{\nu_S^{+,\pi}(s)}u(s,a).$$

In particular, $W_\pi \succeq \alpha_\pi I$. Both summands vanish on $\mathcal{H}_C^\perp$. Eigenvalue monotonicity therefore gives, for $i \le q$,

$$\sigma_i(\mathcal{A}_\pi)^2 \ge \alpha_\pi + \sigma_i(\mathcal{D}_\pi)^2 = \alpha_\pi + \left(\frac{\gamma}{1-\gamma}\right)^2 \left(\sigma_i^{+,\pi}\right)^2.$$

The best rank-k operator-norm approximation error is $\sigma_{k+1}(\mathcal{A}_\pi)^2$, proving the lower bound. The SUAF error follows from truncating the singular expansion of $\mathcal{D}_\pi$.

When $\nu^{+,\pi} = d^\pi$, $W_\pi = I$ on $\mathcal{H}_C$, so all the above eigenvalue inequalities become equalities. □

D.3 Proof of Theorem 3.2

Theorem D.3 (Average-case optimality of the rank-k representations)**.** *Let reward function r be sampled from uniform distribution over the unit sphere of $L^2(\nu^{+,\pi})$, and let $m = |\mathrm{supp}(\nu^{+,\pi})|$. Among all reward-independent rank-k approximations of the positive-lag future operator $\mathcal{T}^{+,\pi}$ in A_r^π, the top-k truncated representation*

$$A_{r,k}^\pi = C_\pi r + \frac{\gamma}{1-\gamma}\sum_{i=1}^{k} \sigma_i^{+,\pi} f_i \langle g_i, r\rangle_{\nu^{+,\pi}}$$

minimizes the expected advantage approximation error, i.e.,

$$\mathbb{E}_r\left[\left\|A_r^\pi - A_{r,k}^\pi\right\|_{d^\pi}^2\right] = \inf_{\mathit{rank}(\widehat{\mathcal{T}}^{+,\pi})\le k} \mathbb{E}_r\left[\left\|A_r^\pi - \hat{A}_r^\pi\right\|_{d^\pi}^2\right] = \frac{1}{m}\left(\frac{\gamma}{1-\gamma}\right)^2 \sum_{i>k}\left(\sigma_i^{+,\pi}\right)^2,$$

and top-k truncation is also optimal under minimax case, i.e.,

$$\sup_{\|r\|_{\nu^{+,\pi}}\le 1}\left\|A_r^\pi - A_{r,k}^\pi\right\|_{d^\pi}^2 = \inf_{\mathit{rank}(\widehat{\mathcal{T}}^{+,\pi})\le k}\sup_{\|r\|_{\nu^{+,\pi}}\le 1}\left\|A_r^\pi - \hat{A}_r^\pi\right\|_{d^\pi}^2 = \left(\frac{\gamma}{1-\gamma}\right)^2\left(\sigma_{k+1}^{+,\pi}\right)^2.$$

Proof. For brevity, let

$$\mathcal{H}_\nu = L^2(\nu^{+,\pi}), \qquad \mathcal{H}_d = L^2(d^\pi),$$

where both spaces are restricted to the supports of their corresponding distributions. Define the coordinate maps

$$[\mathcal{U}_\nu r](s^+,a^+) = \sqrt{\nu^{+,\pi}(s^+,a^+)}\, r(s^+,a^+), \quad [\mathcal{U}_d h](s,a) = \sqrt{d^\pi(s,a)}\, h(s,a).$$

These maps are isometries because

$$\|\mathcal{U}_\nu r\|_2 = \|r\|_{\nu^{+,\pi}}, \qquad \|\mathcal{U}_d h\|_2 = \|h\|_{d^\pi}.$$

By the definitions of $\mathcal{T}^{+,\pi}$ and the positive-lag DCDM, for every $r \in \mathcal{H}_\nu$,

$$[\mathcal{U}_d \mathcal{T}^{+,\pi} r](s,a) = \sum_{s^+,a^+} \widetilde{\boldsymbol{B}}^{+,\pi}\big((s,a),(s^+,a^+)\big)[\mathcal{U}_\nu r](s^+,a^+).$$

Therefore, $\mathcal{T}^{+,\pi}$ and $\widetilde{\boldsymbol{B}}^{+,\pi}$ have the same rank and singular values

$$\mathcal{U}_d \mathcal{T}^{+,\pi} \mathcal{U}_\nu^{-1} = \widetilde{\boldsymbol{B}}^{+,\pi}.$$

In particular, the SVD of the DCDM implies

$$\mathcal{T}^{+,\pi} r = \sum_{i=1}^{R} \sigma_i^{+,\pi} f_i \langle g_i, r \rangle_{\nu^{+,\pi}}.$$

Its rank-k truncation is

$$\mathcal{T}_k^{+,\pi} r = \sum_{i=1}^{k} \sigma_i^{+,\pi} f_i \langle g_i, r \rangle_{\nu^{+,\pi}}.$$

Now let $\widehat{\mathcal{T}}^{+,\pi}$ be any reward-independent operator with $\operatorname{rank}(\widehat{\mathcal{T}}^{+,\pi}) \le k$, and define

$$\widehat{\boldsymbol{B}} = \mathcal{U}_d \widehat{\mathcal{T}}^{+,\pi} \mathcal{U}_\nu^{-1}.$$

Since $\mathcal{U}_d$ and $\mathcal{U}_\nu$ are invertible isometries,

$$\operatorname{rank}(\widehat{\boldsymbol{B}}) = \operatorname{rank}(\widehat{\mathcal{T}}^{+,\pi}) \le k.$$

Moreover, because the immediate term $C_\pi r$ is retained exactly,

$$A_r^\pi - \widehat{A}_r^\pi = \frac{\gamma}{1-\gamma}\left(\mathcal{T}^{+,\pi} - \widehat{\mathcal{T}}^{+,\pi}\right) r.$$

For the average-case result, let

$$\boldsymbol{z}_r = \mathcal{U}_\nu r.$$

Since $\mathcal{U}_\nu$ is an isometry, if r is uniformly distributed over the unit sphere of $\mathcal{H}_\nu$, then $\boldsymbol{z}_r$ is uniformly distributed over the Euclidean unit sphere in $\mathbb{R}^m$. Hence,

$$\mathbb{E}_r\left[\boldsymbol{z}_r \boldsymbol{z}_r^{\mathrm{T}}\right] = \frac{1}{m} \boldsymbol{I}_m.$$

It follows that

$$\begin{aligned}
\mathbb{E}_r\left[\left\|A_r^\pi - \widehat{A}_r^\pi\right\|_{d^\pi}^2\right] &= \left(\frac{\gamma}{1-\gamma}\right)^2 \mathbb{E}_r\left[\left|\left(\widetilde{\boldsymbol{B}}^{+,\pi} - \widehat{\boldsymbol{B}}\right)\boldsymbol{z}_r\right|_2^2\right] \\
&= \left(\frac{\gamma}{1-\gamma}\right)^2 \operatorname{Tr}\left[\left(\widetilde{\boldsymbol{B}}^{+,\pi} - \widehat{\boldsymbol{B}}\right)^{\mathrm{T}}\left(\widetilde{\boldsymbol{B}}^{+,\pi} - \widehat{\boldsymbol{B}}\right)\mathbb{E}_r\left[\boldsymbol{z}_r \boldsymbol{z}_r^{\mathrm{T}}\right]\right] \\
&= \frac{1}{m}\left(\frac{\gamma}{1-\gamma}\right)^2 \left\|\widetilde{\boldsymbol{B}}^{+,\pi} - \widehat{\boldsymbol{B}}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

By the Eckart-Young-Mirsky theorem (Eckart & Young, 1936),

$$\inf_{\operatorname{rank}(\widehat{\boldsymbol{B}}) \le k} \left\|\widetilde{\boldsymbol{B}}^{+,\pi} - \widehat{\boldsymbol{B}}\right\|_{\mathrm{F}}^2 = \sum_{i>k} \left(\sigma_i^{+,\pi}\right)^2,$$

and the infimum is attained by the truncated SVD

$$\widetilde{\boldsymbol{B}}_k^{+,\pi} = \sum_{i=1}^{k} \sigma_i^{+,\pi} \boldsymbol{\psi}_{i,+,\pi}^{S,A} \left(\boldsymbol{\psi}_{i,+,\pi}^{S^+,A^+}\right)^{\mathrm{T}}.$$

This matrix corresponds exactly to $\mathcal{T}_k^{+,\pi}$ and therefore to $A_{r,k}^\pi$. Consequently,

$$\inf_{\operatorname{rank}(\widehat{\mathcal{T}^{+,\pi}}) \le k} \mathbb{E}_r\left[\left\|A_r^\pi - \widehat{A}_r^\pi\right\|_{d^\pi}^2\right] = \frac{1}{m}\left(\frac{\gamma}{1-\gamma}\right)^2 \sum_{i>k} \left(\sigma_i^{+,\pi}\right)^2.$$

For the minimax result, for any $\widehat{\mathcal{T}}^{+,\pi}$ with rank at most k,

$$\begin{aligned}\sup_{\|r\|_*\nu^{+,\pi}\le 1}\left\|A_r^\pi-\widehat{A}_r^\pi\right\|_{d^\pi}^2 &= \left(\frac{\gamma}{1-\gamma}\right)^2 \sup_{\|r\|_{\nu^{+,\pi}}\le 1}\left\|\left(\mathcal{T}^{+,\pi}-\widehat{\mathcal{T}}^{+,\pi}\right)r\right\|_{d^\pi}^2\\ &= \left(\frac{\gamma}{1-\gamma}\right)^2\left\|\mathcal{T}^{+,\pi}-\widehat{\mathcal{T}}^{+,\pi}\right\|_{\mathrm{op}}^2\\ &= \left(\frac{\gamma}{1-\gamma}\right)^2\left\|\widetilde{\boldsymbol{B}}^{+,\pi}-\widehat{\boldsymbol{B}}\right\|_2^2,\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_2$ denote the induced operator norm and the matrix spectral norm, respectively. Applying the Eckart–Young–Mirsky theorem in the spectral norm gives

$$\inf_{\mathrm{rank}(\widehat{\boldsymbol{B}})\le k}\left\|\widetilde{\boldsymbol{B}}^{+,\pi}-\widehat{\boldsymbol{B}}\right\|_2=\sigma_{k+1}^{+,\pi}.$$

The truncated matrix $\widetilde{\boldsymbol{B}}_k^{+,\pi}$ attains this infimum. Hence,

$$\sup_{\|r\|_{\nu^{+,\pi}}\le 1}\left\|A_r^\pi-A_{r,k}^\pi\right\|_{d^\pi}^2=\inf_{\mathrm{rank}(\widehat{\mathcal{T}}^{+,\pi})\le k}\sup_{\|r\|_{\nu^{+,\pi}}\le 1}\left\|A_r^\pi-\widehat{A}_r^\pi\right\|_{d^\pi}^2=\left(\frac{\gamma}{1-\gamma}\right)^2\left(\sigma_{k+1}^{+,\pi}\right)^2.$$

When $k<R$, the worst-case error of the truncated representation is attained, for example, by $r=g_{k+1}$. If $k\ge R$, both the average-case and minimax errors are zero, with the convention $\sigma_{k+1}^{+,\pi}=0$. □

D.4 Proof of Theorem 3.1

Theorem D.4 (Universal advantage function representation)**.** *For any reward function $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$, and for every (s,a) in the support of d^π, the advantage function admits the exact representation*

$$A_r^\pi(s,a)=C_\pi r(s,a)+\frac{\gamma}{1-\gamma}\sum_{i=1}^{R}\sigma_i^{+,\pi}f_i(s,a)\,\langle g_i,r\rangle_{\nu^{+,\pi}},$$

where $C_\pi r(s,a):=r(s,a)-\mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]$ is the immediate action-centered reward and $\langle g_i,r\rangle_{\nu^{+,\pi}}=\sum_{s,a}\nu^{+,\pi}(s,a)g_i(s,a)r(s,a)$ is the inner product on Hilbert space $L^2(\nu^{+,\pi})$. Moreover, for any $k\le R$, define the rank-k approximation

$$A_{r,k}^\pi(s,a):=C_\pi r(s,a)+\frac{\gamma}{1-\gamma}\sum_{i=1}^{k}\sigma_i^{+,\pi}f_i(s,a)\beta_i^{+,\pi}(r).$$

Its approximation error is given exactly by

$$\left\|A_r^\pi-A_{r,k}^\pi\right\|_{L^2(d^\pi)}^2=\left(\frac{\gamma}{1-\gamma}\right)^2\sum_{i=k+1}^{R}\left(\sigma_i^{+,\pi}\right)^2\left(\beta_i^{+,\pi}(r)\right)^2.$$

Proof. The SVD of $\widetilde{\boldsymbol{B}}^{+,\pi}$, together with the definitions of f_i and g_i, gives

$$\begin{aligned}&M^{+,\pi}(s^+,a^+\mid s,a)-\mathbb{E}_{a'\sim\pi(\cdot|s)}\left[M^{+,\pi}(s^+,a^+\mid s,a')\right]\\ &\quad=\nu^{+,\pi}(s^+,a^+)\sum_{i=1}^{R}\sigma_i^{+,\pi}f_i(s,a)g_i(s^+,a^+).\end{aligned}\tag{54}$$

Multiplying both sides by $r(s^+,a^+)$ and summing over (s^+,a^+) yields

$$\sum_{s^+,a^+}\left(M^{+,\pi}(s^+,a^+|s,a)-\mathbb{E}_{a'\sim\pi(\cdot|s)}\left[M^{+,\pi}(s^+,a^+|s,a')\right]\right)r(s^+,a^+)=\sum_{i=1}^{R}\sigma_i^{+,\pi}f_i(s,a)\beta_i^{+,\pi}(r).\tag{55}$$

By the definition of the normalized positive-lag successor measure,

$$\frac{\gamma}{1-\gamma}\sum_{s^+,a^+} M^{+,\pi}(s^+,a^+ \mid s,a)r(s^+,a^+) = \sum_{t=1}^{\infty}\gamma^t \mathbb{E}_\pi\left[r(S_t,A_t) \mid S_0=s, A_0=a\right]. \tag{56}$$

Therefore,

$$Q_r^\pi(s,a) = r(s,a) + \frac{\gamma}{1-\gamma}\sum_{s^+,a^+} M^{+,\pi}(s^+,a^+ \mid s,a)r(s^+,a^+). \tag{57}$$

Averaging this identity over $a' \sim \pi(\cdot \mid s)$ gives $V_r^\pi(s)$. Subtracting $V_r^\pi(s)$ from $Q_r^\pi(s,a)$ and applying equation 55 proves equation 4.

Finally, the functions $\{f_i\}_{i=1}^R$ are orthonormal in $L^2(d^\pi)$:

$$\mathbb{E}_{(S,A)\sim d^\pi}\left[f_i(S,A)f_j(S,A)\right] = \delta_{ij}. \tag{58}$$

Hence the omitted spectral components are mutually orthogonal, and Parseval's identity gives

$$\left\|A_r^\pi - A_{r,k}^\pi\right\|_{L^2(d^\pi)}^2 = \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_{i=k+1}^{R}\left(\sigma_i^{+,\pi}\right)^2\left(\beta_i^{+,\pi}(r)\right)^2. \tag{59}$$

□

D.5 SUAF will Find Optimal Policy in Zero-shot Manner

Conventions and assumptions. Let $n = |\mathcal{S}||\mathcal{A}|$, and represent rewards and normalized discounted occupancies as column vectors in $\mathbb{R}^n$. In particular,

$$d^\pi(s,a) = (1-\gamma)\sum_{t\geq 0}\gamma^t \Pr_{\mu,\pi}(s_t=s, a_t=a), \quad J_r(\pi) = \frac{r^\top d^\pi}{1-\gamma}.$$

Optimality throughout is with respect to the fixed initial distribution μ. At states with zero occupancy, the recovered policy can be chosen arbitrarily. Full support of d_0 and ν makes all weighted feature coordinates well-defined. The statements concern exact SUAF features, exact reference distributions, and exact optimization.

Define the state lifting matrix $H \in \mathbb{R}^{n\times|\mathcal{S}|}$, the reference-policy averaging matrix $\Pi_0 \in \mathbb{R}^{|\mathcal{S}|\times n}$, and the environment transition matrix $T \in \mathbb{R}^{n\times|\mathcal{S}|}$ by

$$(Hw)(s,a) = w(s), \quad (\Pi_0 x)(s) = \sum_a \pi_0(a \mid s)x(s,a), \quad (Tw)(s,a) = \sum_{s'} T(s' \mid s,a)w(s').$$

Thus $\Pi_0 H = I$, $P_0 = T\Pi_0$, and $C_0 = I - H\Pi_0$. The full advantage operator and its delayed component are

$$\mathcal{A}_0 = C_0(I-\gamma P_0)^{-1}, \qquad D_0 = \mathcal{A}_0 - C_0 = cF\Sigma G^\top W_\nu.$$

Here the positive-lag outcome distribution is $\nu^\top = d_0^\top M^{+,\pi_0}$, with $M^{+,\pi_0} = (1-\gamma)P_0(I-\gamma P_0)^{-1}$. The feature normalizations are

$$F^\top W_d F = I_R, \qquad G^\top W_\nu G = I_R, \qquad C_0 F = F.$$

All transposes are ordinary coordinate transposes.

Flow constraints and policy recovery. Set $B = H^\top - \gamma T^\top$. The feasible occupancy set is

$$\mathcal{D}_\mu = \{d \geq 0 : Bd = (1-\gamma)\mu\}. \tag{60}$$

Necessity follows from the discounted occupancy identity. For sufficiency, given such a d, put $d_S(s) = \sum_a d(s,a)$ and define $\pi_d(a \mid s) = d(s,a)/d_S(s)$ whenever $d_S(s) > 0$. At a zero-mass state choose any action distribution. Then $d(s,a) = d_S(s)\pi_d(a \mid s)$ everywhere, and the flow constraint becomes

$$d_S = (1-\gamma)\mu + \gamma(P_S^{\pi_d})^\top d_S.$$

Its unique solution is the normalized discounted state occupancy of π_d, since $I - \gamma(P_S^{\pi_d})^\top$ is invertible. Thus $d = d^{\pi_d}$. Summing the flow constraints also gives $\mathbf{1}^\top d = 1$, so $\mathcal{D}_\mu$ is compact.

A simple occupancy-flow identity makes SUAF sufficient for control. In a finite MDP with $0 < \gamma < 1$, fix μ and π_0 with full-support $d_0 = d^{\pi_0}$ and $\nu = \nu^{+,\pi_0}$. Let F, G, Σ collect all R SUAF modes, and write $W_d = \mathrm{diag}(d_0)$, $W_\nu = \mathrm{diag}(\nu)$, and $c = \gamma/(1-\gamma)$.

Theorem D.5 (Policy recovery)**.** *Let Z be a given basis matrix for the subspace with $\sum_a z(s,a) = 0$ for every s, with $q = |\mathcal{S}|(|\mathcal{A}| - 1)$ columns. Define*

$$\Phi = Z + cW_\nu G\Sigma F^\top Z, \quad \mathcal{D}_\mu = \{d_0 + \Phi w : \ w \in \mathbb{R}^q, \ d_0 + \Phi w \geq 0\}, \tag{61}$$

then $\mathcal{D}_\mu$ is the set of all stationary-policy occupancies. For any reward r, maximizing $r^\top(d_0 + \Phi w)$ over this set and normalizing the optimizer statewise recovers a μ-optimal policy.

The matrix Z is prescribed, with correction having rank at most $R \leq |\mathcal{S}| - 1$. SUAF requires no further environment interaction. With fewer modes, we restrict search to feasible directions.

Proof of Theorem D.5. The elementary identities needed for the construction are

$$B\mathcal{A}_0^\top = 0, \qquad \mathcal{A}_0 C_0 = C_0. \tag{62}$$

Indeed, $P_0 H = T$, and hence $B^\top = H - \gamma T = (I - \gamma P_0)H$. Consequently, $\mathcal{A}_0 B^\top = C_0 H = 0$. Also, $P_0 C_0 = T\Pi_0 C_0 = 0$, so $(I - \gamma P_0)^{-1} C_0 = C_0$ and $\mathcal{A}_0 C_0 = C_0^2 = C_0$.

Because H has full column rank and $I - \gamma P_0$ is invertible, B has rank $|\mathcal{S}|$. Therefore $\dim\ker B = q$. The columns of Z span $\ker H^\top$, so $C_0^\top Z = (I - \Pi_0^\top H^\top)Z = Z$. It follows that

$$\mathcal{A}_0^\top Z = C_0^\top Z + D_0^\top Z = Z + cW_\nu G\Sigma F^\top Z = \Phi.$$

equation 62 yields $B\Phi = 0$. Moreover,

$$C_0^\top \Phi = C_0^\top \mathcal{A}_0^\top Z = (\mathcal{A}_0 C_0)^\top Z = Z.$$

Thus $\mathrm{rank}(\Phi) = q$, since Z has rank q. We conclude that $\mathrm{col}(\Phi) = \ker B$. As $Bd_0 = (1 - \gamma)\mu$, every solution of the flow equalities is uniquely expressible as $d_0 + \Phi w$. Intersecting with the nonnegative orthant proves equation 61.

The reward objective satisfies $r^\top d = (1-\gamma)J_r(\pi_d)$. Optimization over the displayed representation is therefore exactly optimization over all stationary-policy occupancies. The maximum is attained by compactness, and statewise normalization recovers a μ-optimal stationary policy. In a finite discounted MDP, stationary policies suffice for optimality. □

Why the learned correction has rank at most $|\mathcal{S}| - 1$. Using the resolvent identity,

$$D_0 = \gamma C_0 T(I - \gamma\Pi_0 T)^{-1}\Pi_0.$$

Since $T\mathbf{1}_{|\mathcal{S}|} = \mathbf{1}_n$ and $C_0\mathbf{1}_n = 0$, the matrix $C_0 T$ has rank at most $|\mathcal{S}| - 1$. Hence

$$R = \mathrm{rank}(D_0) \leq |\mathcal{S}| - 1, \qquad \mathrm{rank}(\Phi - Z) \leq R.$$

For example, Z can be constructed from columns $e_{(s,a)} - e_{(s,a_0)}$, one for each state and each non-reference action. Thus Z requires no transition-model estimation. This rank bound concerns the learned correction, not the q free coordinates of the full occupancy polytope.

Corollary D.6 (Nested feasible policy search)**.** *Under the same assumptions, define $u_i = W_d f_i + c\sigma_i W_\nu g_i$ and $U_k = [u_1, \ldots, u_k]$. Every nonnegative $d_0 + U_k\alpha$ is a valid occupancy. Let $\pi_{r,k}$ be recovered from an exact solution of $\max_{\alpha:\, d_0 + U_k\alpha \geq 0} r^\top(d_0 + U_k\alpha)$. Then, for $0 \leq k < R$,*

$$J_r(\pi_0) \leq J_r(\pi_{r,k}) \leq J_r(\pi_{r,k+1}) \leq \max_\pi J_r(\pi). \tag{63}$$

However, the improvement is non-strict and retaining all R directions alone need not span $\mathcal{D}_\mu$.

Proof of Corollary D.6. Because $d_0(s,a) = d_{0,S}(s)\pi_0(a \mid s)$, action centering is self-adjoint in $L^2(d_0)$: $C_0^\top W_d = W_d C_0$. For each exact SUAF mode, $C_0 f_i = f_i$ and $F^\top W_d f_i = e_i$. Therefore

$$\begin{aligned}\mathcal{A}_0^\top W_d f_i &= C_0^\top W_d f_i + cW_\nu G\Sigma F^\top W_d f_i \\ &= W_d f_i + c\sigma_i W_\nu g_i = u_i.\end{aligned}$$

equation 62 gives $Bu_i = 0$. Consequently, $d_0 + U_k\alpha$ always satisfies the exact flow equalities and is an occupancy if and only if it is nonnegative.

Define

$$\mathcal{D}_k = (d_0 + \mathrm{col}(U_k)) \cap \mathbb{R}^n_+, \qquad \mathcal{D}_0 = \{d_0\}.$$

These are nonempty compact subsets of $\mathcal{D}_\mu$, and

$$\{d_0\} = \mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \cdots \subseteq \mathcal{D}_R \subseteq \mathcal{D}_\mu.$$

Maximizing the same linear reward objective over these nested sets, and then dividing by $1-\gamma$, proves equation 63. No claim of strict improvement or of optimal mode selection for control is required. □

Completion to a globally sufficient search space. The R nonzero future modes need not supply all q occupancy directions. Define

$$\mathcal{N} = \ker H^\top \cap \ker D_0^\top.$$

For $z \in \mathcal{N}$, $C_0^\top z = z$ and $\mathcal{A}_0^\top z = z$, so $z \in \ker B$. Since $D_0^\top = cW_\nu G \Sigma F^\top$ has rank R and $F^\top W_d F = I_R$, its restriction to $\ker H^\top$ also has rank R. Indeed, each $W_d f_i$ belongs to $\ker H^\top$. Consequently $\dim \mathcal{N} = q - R$ and

$$\ker H^\top = \mathrm{span}\{W_d f_i\}_{i=1}^R \oplus \mathcal{N}.$$

The map $\mathcal{A}_0^\top$ is a bijection from $\ker H^\top$ to $\ker B$, as follows from the proof of the theorem. Applying it to the preceding decomposition gives

$$\ker B = \mathrm{span}\{u_i\}_{i=1}^R \oplus \mathcal{N}.$$

Thus adding a basis of $\mathcal{N}$ makes the search globally sufficient at $k = R$. Identifying this completion generally requires the full exact representation. It is not implied by knowing only the first $k < R$ pairs. Alternatively, the full construction with Z in Theorem D.5 supplies all directions directly.

Scope. These are exact SUAFs. With learned features, errors can break the identity $Bu_i = 0$, and nonnegativity alone no longer certifies feasibility in the true MDP. Furthermore, monotonicity in k follows from nesting, not from a theorem that leading singular modes minimize control regret. The construction does not assert that R is the dimension of the entire policy space or that it is directly comparable to the number of basis matrices in a successor-measure representation.

D.6 One step update with SUAF

Lemma D.7 (Boltzmann update satisfies the improvement condition)**.** *Let π be a behavior policy, and let $A^\pi_{r,k}$ be an advantage score satisfying the centering condition*

$$\mathbb{E}_{a\sim\pi(\cdot|s)}\big[A^\pi_{r,k}(s,a)\big] = 0 \qquad \forall s. \tag{64}$$

For any inverse temperature $\tau > 0$, define the Boltzmann update

$$\pi_r^+(a \mid s) = \frac{\pi(a \mid s)\exp\big(\tau A^\pi_{r,k}(s,a)\big)}{Z^\pi(s)}, \qquad Z^\pi(s) = \int \pi(a' \mid s)\exp\big(\tau A^\pi_{r,k}(s,a')\big)\,da'. \tag{65}$$

Then

$$\mathbb{E}_{a\sim\pi_r^+(\cdot|s)}\big[A^\pi_{r,k}(s,a)\big] \geq 0 \qquad \forall s. \tag{66}$$

Proof. Fix s. Write $\mathbb{E}_\pi[\cdot] = \mathbb{E}_{a\sim\pi(\cdot|s)}[\cdot]$ and $A = A^\pi_{r,k}(s,a)$. From equation 65,

$$\pi_r^+(a \mid s) = \frac{\pi(a \mid s)e^{\tau A}}{Z^\pi(s)}, \qquad Z^\pi(s) = \mathbb{E}_\pi[e^{\tau A}].$$

Thus

$$\mathbb{E}_{\pi_r^+}[A] = \frac{\mathbb{E}_\pi[Ae^{\tau A}]}{\mathbb{E}_\pi[e^{\tau A}]}.$$

Equivalently, using the KL identity,

$$\tau\,\mathbb{E}_{\pi_r^+}[A] = \mathrm{KL}\big(\pi_r^+(\cdot \mid s)\,\|\,\pi(\cdot \mid s)\big) + \log \mathbb{E}_\pi[e^{\tau A}].$$

Since $\log$ is concave, Jensen's inequality gives

$$\log \mathbb{E}_\pi[e^{\tau A}] \geq \tau\, \mathbb{E}_\pi[A] = 0,$$

where the last equality follows from equation 64. Hence

$$\tau\, \mathbb{E}_{\pi_r^+}[A] \geq \mathrm{KL}\big(\pi_r^+(\cdot \mid s) \,\|\, \pi(\cdot \mid s)\big) \geq 0.$$

Dividing by $\tau > 0$ yields equation 66. □

Remark (Automatic centering in the SUAF construction). In our construction,

$$A_{r,k}^\pi = C_\pi r + \frac{\gamma}{1-\gamma} \sum_{i=1}^{k} \sigma_i^{+,\pi} f_i \langle g_i, r \rangle_{\nu^{+,\pi}}.$$

By definition of C_π, the term $C_\pi r$ is centered under π. Moreover, the property $f_i = C_\pi f_i$ ensures that the remaining term is also centered under π. Therefore

$$\mathbb{E}_{a\sim\pi(\cdot|s)}\big[A_{r,k}^\pi(s,a)\big] = 0 \qquad \forall s.$$

Hence the centering condition equation 64 in Lemma D.7 is automatically satisfied, and the Boltzmann update equation 65 satisfies the improvement condition equation 66.

Theorem D.8 (Policy improvement with truncated SUAF)**.** *Consider a policy π_r^+ using Boltzmann update and greedy update. Assume $d^{\pi_r^+} \ll d^\pi$ and define $\chi_k^2 := \mathbb{E}_{d^\pi}[(\frac{d^{\pi_r^+}}{d^\pi} - 1)^2] < \infty$ and $\mathcal{I}_k(r) := \mathbb{E}_{d^{\pi_r^+}}[A_{r,k}^\pi]/(1-\gamma)$. Then $\mathcal{I}_k(r) \geq 0$ and*

$$J_r(\pi_r^+) - J_r(\pi) \geq \mathcal{I}_k(r) - \frac{\gamma\sqrt{\chi_k^2}}{(1-\gamma)^2} \left[\sum_{i>k} (\sigma_i^{+,\pi})^2 \beta_i(r)^2\right]^{1/2}.$$

Consequently, policy improvement is guaranteed whenever the predicted gain exceeds the spectral-tail penalty. For $k = R$, the penalty vanishes.

Proof. By Lemma D.7, we have that $\mathbb{E}_{\pi_r^+}[A_{r,k}^\pi(s,a)] \geq 0$ for every s. Write $e_{r,k} = A_r^\pi - A_{r,k}^\pi$. The performance difference identity gives

$$J_r(\pi_r^+) - J_r(\pi) = \mathcal{I}_k(r) + \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_r^+}}[e_{r,k}].$$

Both A_r^π and $A_{r,k}^\pi$ are action-centered under π, so $\mathbb{E}_{d^\pi}[e_{r,k}] = 0$. Therefore, Cauchy–Schwarz yields

$$\left|\mathbb{E}_{d^{\pi_r^+}}[e_{r,k}]\right| = \left|\mathbb{E}_{d^\pi}\left[\left(\frac{d^{\pi_r^+}}{d^\pi} - 1\right) e_{r,k}\right]\right| \leq \sqrt{\chi_k^2}\, \|e_{r,k}\|_{d^\pi}.$$

By orthonormality of the left singular functions,

$$\|e_{r,k}\|_{d^\pi}^2 = \left(\frac{\gamma}{1-\gamma}\right)^2 \sum_{i>k} (\sigma_i^{+,\pi})^2 \beta_i(r)^2.$$

Combining these identities proves the claim. □

Actually, we have a stronger version of this theorem as Theorem D.9. We write it formal as following:

Theorem D.9 (Policy improvement with truncated SUAF Formal)**.** *Let $h = A_{r,k}^\pi$ and $e = A_r^\pi - h$. Let π_r^+ be obtained either by a greedy update with respect to h, or by the Boltzmann update*

$$\pi_\tau(a \mid s) = \frac{\pi(a \mid s) \exp(\tau h(s,a))}{\mathbb{E}_{a'\sim\pi(\cdot|s)}[\exp(\tau h(s,a'))]}, \qquad \tau > 0.$$

Assume $d^{\pi_r^+} \ll d^\pi$ and $w := d^{\pi_r^+}/d^\pi \in L^2(d^\pi)$. Define

$$\bar{w}(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[w(s,a)], \qquad \chi_{k,\mathrm{act}}^2 := \mathbb{E}_{d^\pi}\big[(w(s,a) - \bar{w}(s))^2\big],$$

and $\mathcal{I}_k(r) := \mathbb{E}_{d^{\pi_r^+}}[h]/(1-\gamma)$. *Then* $\mathcal{I}_k(r) \geq 0$ *and*

$$J_r(\pi_r^+) - J_r(\pi) \geq \mathcal{I}_k(r) - \frac{\gamma\sqrt{\chi^2_{k,\mathrm{act}}}}{(1-\gamma)^2}\left[\sum_{i>k}(\sigma_i^{+,\pi})^2\beta_i(r)^2\right]^{1/2}.$$

Moreover, $\chi^2_{k,\mathrm{act}} \leq \chi^2_k := \mathbb{E}_{d^\pi}[(w-1)^2]$. *Thus the bound is at least as tight as the bound using* χ^2_k, *and its penalty vanishes when* $k = R$.

For the Boltzmann update, write $\pi_r^+ = \pi_\tau$ *and suppose*

$$B := \sup_{s,a,b}|h(s,a) - h(s,b)| < \infty, \qquad D := \sup_{s,a,b}|A_r^\pi(s,a) - A_r^\pi(s,b)| < \infty.$$

With $m := \langle h, A_r^\pi\rangle_{d^\pi}$, *we additionally have*

$$J_r(\pi_\tau) - J_r(\pi) \geq \frac{\tau}{1-\gamma}\left(m - \frac{DB^2}{8(1-\gamma)}\tau\right).$$

Consequently, if $m > 0$, *every*

$$0 < \tau < \frac{8(1-\gamma)m}{DB^2}$$

guarantees strict policy improvement. In particular,

$$m = \|h\|^2_{d^\pi} + \langle C^\pi r, e\rangle_{d^\pi} \geq \|h\|^2_{d^\pi} - \|C^\pi r\|_{d^\pi}\|e\|_{d^\pi}.$$

For state-only rewards, $C^\pi r = 0$ *and hence* $m = \|h\|^2_{d^\pi} > 0$ *whenever* $h \neq 0$.

Proof. Both h and A_r^π are action-centered under π, so $\mathbb{E}_{a\sim\pi(\cdot|s)}h(s,a) = \mathbb{E}_{a\sim\pi(\cdot|s)}e(s,a) = 0$. A greedy update satisfies $\mathbb{E}_{\pi_r^+}[h(s,a) \mid s] = \max_a h(s,a) \geq 0$. For the Boltzmann update, let $Z_\tau(s) = \mathbb{E}_\pi[\exp(\tau h(s,a)) \mid s]$. Then

$$\tau\mathbb{E}_{\pi_\tau}[h(s,a) \mid s] = D_{\mathrm{KL}}\big(\pi_\tau(\cdot \mid s)\|\pi(\cdot \mid s)\big) + \log Z_\tau(s) \geq 0,$$

where $\log Z_\tau(s) \geq \tau\mathbb{E}_\pi[h(s,a) \mid s] = 0$ by Jensen's inequality. Thus $\mathcal{I}_k(r) \geq 0$ in either case.

The performance-difference identity and the SUAF expansion give

$$J_r(\pi_r^+) - J_r(\pi) = \mathcal{I}_k(r) + \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_r^+}}[e],$$
$$e = \frac{\gamma}{1-\gamma}\sum_{i>k}\sigma_i^{+,\pi}\beta_i(r)f_i,$$
$$\|e\|^2_{d^\pi} = \left(\frac{\gamma}{1-\gamma}\right)^2\sum_{i>k}(\sigma_i^{+,\pi})^2\beta_i(r)^2.$$

Since e is conditionally action-centered,

$$\mathbb{E}_{d^{\pi_r^+}}[e] = \mathbb{E}_{d^\pi}[we] = \mathbb{E}_{d^\pi}[(w-\bar{w})e].$$

Cauchy–Schwarz proves the first bound. Conditional orthogonality also gives

$$\chi^2_k = \mathbb{E}_{d^\pi}[(w-\bar{w})^2] + \mathbb{E}_{s\sim d_S^\pi}[(\bar{w}(s)-1)^2] \geq \chi^2_{k,\mathrm{act}}.$$

It remains to prove the temperature bound. For each state, define $F_s(t) = \mathbb{E}_{a\sim\pi_t(\cdot|s)}[A_r^\pi(s,a)]$, so that $F_s(0) = 0$. Differentiating the Boltzmann distribution yields

$$F_s'(t) = \mathrm{Cov}_{\pi_t}(A_r^\pi(s,\cdot), h(s,\cdot)).$$

The covariance bound for bounded random variables implies $|F_s'(t)| \leq DB/4$. Differentiating once more gives

$$F_s''(t) = \mathrm{Cov}_{\pi_t}\left(A_r^\pi(s,\cdot), \big(h(s,\cdot) - \mathbb{E}_{\pi_t}h(s,\cdot)\big)^2\right), \qquad |F_s''(t)| \leq \frac{DB^2}{4}.$$

Consequently,

$$F_s(\tau) \geq \tau F_s'(0) - \frac{DB^2}{8}\tau^2, \qquad |F_s(\tau)| \leq \frac{DB}{4}\tau.$$

The same covariance bound applied to action-set indicators gives $\|\pi_\tau(\cdot \mid s) - \pi(\cdot \mid s)\|_1 \leq \tau B/2$. The discounted occupancy equations therefore imply

$$\|d_S^{\pi_\tau} - d_S^\pi\|_1 \leq \frac{\gamma\tau B}{2(1-\gamma)}.$$

Since $m = \mathbb{E}_{s\sim d_S^\pi}[F_s'(0)]$, combining the preceding bounds yields

$$\begin{aligned}\mathbb{E}_{s\sim d_S^{\pi_\tau}}[F_s(\tau)] &\geq \mathbb{E}_{s\sim d_S^\pi}[F_s(\tau)] - \|d_S^{\pi_\tau} - d_S^\pi\|_1 \sup_s |F_s(\tau)| \\ &\geq \tau m - \frac{DB^2}{8(1-\gamma)}\tau^2.\end{aligned}$$

The performance-difference identity now proves the stated temperature bound.

Finally, the retained SUAF modes are orthogonal in $L^2(d^\pi)$ to the discarded modes. Since $h = C^\pi r$ plus the retained modes and e consists of discarded modes, $\langle h, e\rangle_{d^\pi} = \langle C^\pi r, e\rangle_{d^\pi}$. This proves the expression and lower bound for m. □

Here, we are ready to prove Theorem 4.1.

Theorem D.10 (Boltzmann policy improvement with truncated SUAF)**.** *Let $A_{r,k}^\pi$ be the exact rank-k SUAF advantage approximation, and let $\pi_\tau(a|s) \propto \pi(a|s)e^{\tau h(s,a)}$ with $\tau > 0$. Define $B_k := \sup_{s,a,b} |A_{r,k}^\pi(s,a) - A_{r,k}^\pi(s,b)|$, $\varepsilon_k(r) = \frac{\gamma}{1-\gamma}\left[\sum_{i>k}(\sigma_i^{+,\pi})^2\beta_i(r)^2\right]^{1/2}$ and $\Delta_k(r) := \|A_{r,k}^\pi\|_{d^\pi}^2 - \|C^\pi r\|_{d^\pi}\varepsilon_k(r)$ If $\Delta_k(r) > 0$, then it guarantees $J_r(\pi_\tau) > J_r(\pi)$ if temperature satisfies $0 < \tau < \frac{4(1-\gamma)^2\Delta_k(r)}{R_{\max}B_k^2}$. For state-only rewards, $C^\pi r = 0$ and $\Delta_k(r) = \left(\frac{\gamma}{1-\gamma}\right)^2\sum_{i\leq k}(\sigma_i^{+,\pi})^2\beta_i(r)^2$. In this case, if $\Delta_k(r) > 0$, the larger sufficient range*

$$0 < \tau < \frac{4\gamma}{R_{\max}B_k^2}\sum_{i\leq k}(\sigma_i^{+,\pi})^2\beta_i(r)^2$$

also guarantees strict improvement.

Proof. Write $h = A_{r,k}^\pi$, $e = A_r^\pi - h$, and $m = \langle h, A_r^\pi\rangle_{d^\pi}$. By orthogonality of the retained and discarded SUAF modes,

$$m = \|h\|_{d^\pi}^2 + \langle C^\pi r, e\rangle_{d^\pi} \geq \|h\|_{d^\pi}^2 - \|C^\pi r\|_{d^\pi}\|e\|_{d^\pi} = \Delta_k(r),$$

where $\|e\|_{d^\pi} = \varepsilon_k(r)$.

Let $D = \sup_{s,a,b}|A_r^\pi(s,a) - A_r^\pi(s,b)|$. Since $|r(s,a)| \leq R_{\max}$, $|Q_r^\pi(s,a)| \leq R_{\max}/(1-\gamma)$. Actions at the same state share the same value baseline; hence

$$D = \sup_{s,a,b}|Q_r^\pi(s,a) - Q_r^\pi(s,b)| \leq \frac{2R_{\max}}{1-\gamma}.$$

Applying the previously established temperature bound gives

$$\begin{aligned}J_r(\pi_\tau) - J_r(\pi) &\geq \frac{\tau}{1-\gamma}\left(m - \frac{DB_k^2}{8(1-\gamma)}\tau\right) \\ &\geq \frac{\tau}{1-\gamma}\left(\Delta_k(r) - \frac{R_{\max}B_k^2}{4(1-\gamma)^2}\tau\right).\end{aligned}$$

The right-hand side is strictly positive under the first stated condition on τ.

If $r = r(s)$, then $C^\pi r = 0$. Orthonormality of the SUAF left features therefore gives

$$\Delta_k(r) = \|h\|_{d^\pi}^2 = \left(\frac{\gamma}{1-\gamma}\right)^2\sum_{i\leq k}(\sigma_i^{+,\pi})^2\beta_i(r)^2.$$

Furthermore, $Q_r^\pi(s,a) = r(s) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_r^\pi(s')]$. The immediate reward cancels when comparing actions at the same state, so

$$D \leq \frac{2\gamma R_{\max}}{1-\gamma}.$$

The temperature bound now yields

$$J_r(\pi_\tau) - J_r(\pi) \geq \frac{\tau}{1-\gamma}\left(\Delta_k(r) - \frac{\gamma R_{\max} B_k^2}{4(1-\gamma)^2}\tau\right).$$

Substituting the expression for $\Delta_k(r)$ shows that this quantity is strictly positive under the second stated condition on τ. □

D.7 Estimation Error of Advantage Function Under Trust Region

For the trust region algorithms, our estimation of advantage function can be bounded. Furthermore, we claim that trust-region-controlled SUAF reuse yields reward-uniform policy-gradient error bounds, with reliable ascent whenever the gradient signal exceeds the reuse error.

Lemma D.11 (Reward-independent transfer bound)**.** *Denote $\Delta A_r^{\pi'} = A_r^{\pi'} - \widehat{A}_r$. Under the above setting, the approximation in equation 7 satisfies*

$$\sup_{\|r\|_{\nu^+,\pi} \leq 1} \left\|\Delta A_r^{\pi'}\right\|_{d^{\pi'}}^2 \leq \frac{\gamma L_{\max}^2}{2(1-\gamma)} \mathcal{D}_{\mathrm{KL}}(\pi' \| \pi), \sup_{\|r\|_{\nu^+,\pi} \leq 1} \left\|\Delta A_r^{\pi'}\right\|_{d^{\pi}}^2 \leq \frac{(4-3\gamma) L_{\max}^2}{2(1-\gamma)} \mathcal{D}_{\mathrm{KL}}(\pi \| \pi')$$

Proof. Fix a reward r with $\|r\|_{\nu^+,\pi} \leq 1$. Write $d_S = d_S^{\pi'}$, and define

$$\Delta V_r(s) := V_r^{\pi'}(s) - V_r^\pi(s), \qquad q_r(s) := \mathbb{E}_{a \sim \pi'(\cdot|s)}[A_r^\pi(s,a)].$$

Throughout the proof, $\|u\|_{d_S}^2 = \sum_s d_S(s) u^2(s)$. Let $P_S^{\pi'}$ denote the state transition operator:

$$(P_S^{\pi'} u)(s) = \sum_{a,s'} \pi'(a \mid s) T(s' \mid s, a) u(s').$$

Step 1: Expressing the transfer error. The Bellman equations imply

$$\Delta V_r = q_r + \gamma P_S^{\pi'} \Delta V_r, \tag{67}$$

in that $q_r(s) = \mathbb{E}_{a \sim \pi'}[Q_r^\pi(s,a)] - V_r^\pi(s)$. Moreover, since $\widehat{A}_r = C_{\pi'} Q_r^\pi$ and $A_r^{\pi'} = C_{\pi'} Q_r^{\pi'}$, we have

$$A_r^{\pi'} - \widehat{A}_r = \gamma C_{\pi'} h_r, \qquad h_r(s,a) := \sum_{s'} T(s' \mid s, a) \Delta V_r(s').$$

Therefore, using the conditional variance identity and Jensen's inequality,

$$\begin{aligned} \|A_r^{\pi'} - \widehat{A}_r\|_{d^{\pi'}}^2 &= \gamma^2 \left(\mathbb{E}_{(s,a) \sim d^{\pi'}}[h_r(s,a)^2] - \|P_S^{\pi'} \Delta V_r\|_{d_S}^2\right) \\ &\leq \gamma^2 \left(\mathbb{E}_{s \sim d_S}[P_S^{\pi'}(\Delta V_r^2)(s)] - \|P_S^{\pi'} \Delta V_r\|_{d_S}^2\right). \end{aligned} \tag{68}$$

Step 2: Using the discounted occupancy identity. The state occupancy satisfies

$$d_S = (1-\gamma)\mu + \gamma d_S P_S^{\pi'}.$$

Consequently,

$$\gamma\, \mathbb{E}_{s \sim d_S}[P_S^{\pi'}(\Delta V_r^2)(s)] = \|\Delta V_r\|_{d_S}^2 - (1-\gamma)\|\Delta V_r\|_\mu^2 \leq \|\Delta V_r\|_{d_S}^2.$$

Combining this inequality with equation 68 and $\gamma P_S^{\pi'} \Delta V_r = \Delta V_r - q_r$ yields

$$\begin{aligned} \|A_r^{\pi'} - \widehat{A}_r\|_{d^{\pi'}}^2 &\leq \gamma \|\Delta V_r\|_{d_S}^2 - \|\Delta V_r - q_r\|_{d_S}^2 \\ &= -(1-\gamma)\left\|\Delta V_r - \frac{q_r}{1-\gamma}\right\|_{d_S}^2 + \frac{\gamma}{1-\gamma}\|q_r\|_{d_S}^2 \\ &\leq \frac{\gamma}{1-\gamma}\|q_r\|_{d_S}^2. \end{aligned} \tag{69}$$

Step 3: Bounding the policy perturbation by KL. Since $\mathbb{E}_{a\sim\pi(\cdot|s)}[A_r^\pi(s,a)]=0$,

$$q_r(s)=\sum_a\big(\pi'(a\mid s)-\pi(a\mid s)\big)A_r^\pi(s,a).$$

For any two probability distributions, their difference in the expectation of a function is bounded by the function's range times their total variation distance. Hence, by the definition of $L_{\max}$,

$$|q_r(s)|\le L_{\max}\ \mathrm{TV}\big(\pi'(\cdot\mid s),\pi(\cdot\mid s)\big),$$

where $\mathrm{TV}(p,q)=\frac{1}{2}\sum_a|p(a)-q(a)|$. Pinsker's inequality (Brillinger, 1964), with natural logarithms, therefore gives

$$\|q_r\|_{d_S}^2\le\frac{L_{\max}^2}{2}\mathbb{E}_{s\sim d_S^{\pi'}}\left[D_{\mathrm{KL}}\big(\pi'(\cdot\mid s)\|\pi(\cdot\mid s)\big)\right].$$

Substituting into equation 69 and taking the supremum over $\|r\|_{\nu^+,\pi}\le 1$ proves

$$\sup_{\|r\|_{\nu^+,\pi}\le1}\|A_r^{\pi'}-\widehat{A}_r\|_{d^{\pi'}}^2\le\frac{\gamma L_{\max}^2}{2(1-\gamma)}\mathbb{E}_{s\sim d_S^{\pi'}}\left[D_{\mathrm{KL}}\big(\pi'(\cdot\mid s)\|\pi(\cdot\mid s)\big)\right].$$

Step 4: Controlling the reference-centered error. We next establish the bound under d^π. Define

$$\Delta Q_r:=Q_r^{\pi'}-Q_r^\pi=\gamma h_r,\qquad \widetilde{q}_r(s):=\mathbb{E}_{a\sim\pi(\cdot|s)}[A_r^{\pi'}(s,a)].$$

The Bellman equations also give

$$\Delta V_r=-\widetilde{q}_r+\gamma P_S^\pi\Delta V_r,$$

where P_S^π is the state transition operator under π. Indeed,

$$\widetilde{q}_r=\mathbb{E}_{a\sim\pi}[Q_r^{\pi'}]-V_r^{\pi'}=\gamma P_S^\pi\Delta V_r-\Delta V_r.$$

Applying the conditional variance identity and Jensen's inequality under the reference policy yields

$$\|C_\pi\Delta Q_r\|_{d^\pi}^2=\gamma^2\left(\mathbb{E}_{d^\pi}[h_r^2]-\|P_S^\pi\Delta V_r\|_{d_S^\pi}^2\right)\le\gamma^2\left(\mathbb{E}_{d_S^\pi}[P_S^\pi(\Delta V_r^2)]-\|P_S^\pi\Delta V_r\|_{d_S^\pi}^2\right).$$

Using $d_S^\pi=(1-\gamma)\mu+\gamma d_S^\pi P_S^\pi$ and $\gamma P_S^\pi\Delta V_r=\Delta V_r+\widetilde{q}_r$, we obtain

$$\begin{aligned}\|C_\pi\Delta Q_r\|_{d^\pi}^2&\le\gamma\|\Delta V_r\|_{d_S^\pi}^2-\|\Delta V_r+\widetilde{q}_r\|_{d_S^\pi}^2\\&=-(1-\gamma)\left\|\Delta V_r+\frac{\widetilde{q}_r}{1-\gamma}\right\|_{d_S^\pi}^2+\frac{\gamma}{1-\gamma}\|\widetilde{q}_r\|_{d_S^\pi}^2\\&\le\frac{\gamma}{1-\gamma}\|\widetilde{q}_r\|_{d_S^\pi}^2.\end{aligned}\tag{70}$$

Since $\mathbb{E}_{a\sim\pi'}[A_r^{\pi'}(s,a)]=0$,

$$\widetilde{q}_r(s)=\sum_a\big(\pi(a\mid s)-\pi'(a\mid s)\big)A_r^{\pi'}(s,a).$$

The definition of $L_{\max}$ and Pinsker's inequality imply

$$\|\widetilde{q}_r\|_{d_S^\pi}^2\le\frac{L_{\max}^2}{2}\mathbb{E}_{s\sim d_S^\pi}\left[D_{\mathrm{KL}}\big(\pi(\cdot\mid s)\|\pi'(\cdot\mid s)\big)\right].$$

Consequently,

$$\|C_\pi\Delta Q_r\|_{d^\pi}^2\le\frac{\gamma L_{\max}^2}{2(1-\gamma)}\mathcal{D}_{\mathrm{KL}}(\pi\|\pi').\tag{71}$$

Step 5: Accounting for the change of centering. The actual transfer error is $C_{\pi'}\Delta Q_r$. Define the state-dependent discrepancy

$$m_r(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[\Delta Q_r(s,a)] - \mathbb{E}_{a\sim\pi'(\cdot|s)}[\Delta Q_r(s,a)].$$

Then

$$A_r^{\pi'} - \widehat{A}_r = C_\pi \Delta Q_r + m_r.$$

Since $C_\pi \Delta Q_r$ has zero conditional action mean under π, it is orthogonal to m_r in $L^2(d^\pi)$. Therefore,

$$\|A_r^{\pi'} - \widehat{A}_r\|_{d^\pi}^2 = \|C_\pi \Delta Q_r\|_{d^\pi}^2 + \|m_r\|_{d_S^\pi}^2. \tag{72}$$

Subtracting a state-dependent baseline does not change the action span. Hence, for every s, a, b,

$$\begin{aligned}|\Delta Q_r(s,a) - \Delta Q_r(s,b)| &\le |A_r^{\pi'}(s,a) - A_r^{\pi'}(s,b)| + |A_r^{\pi}(s,a) - A_r^{\pi}(s,b)| \\ &\le 2L_{\max}.\end{aligned}$$

It follows that

$$|m_r(s)| \le 2L_{\max}\ \mathrm{TV}\big(\pi(\cdot \mid s), \pi'(\cdot \mid s)\big).$$

Applying Pinsker's inequality gives

$$\|m_r\|_{d_S^\pi}^2 \le 2L_{\max}^2\, \mathcal{D}_{\mathrm{KL}}(\pi\|\pi').$$

Combining this inequality with equation 71 and equation 72 yields

$$\begin{aligned}\|A_r^{\pi'} - \widehat{A}_r\|_{d^\pi}^2 &\le \left(\frac{\gamma}{2(1-\gamma)} + 2\right) L_{\max}^2\, \mathcal{D}_{\mathrm{KL}}(\pi\|\pi') \\ &= \frac{(4-3\gamma)L_{\max}^2}{2(1-\gamma)} \mathcal{D}_{\mathrm{KL}}(\pi\|\pi').\end{aligned}$$

Finally, taking the supremum over the same reward class gives

$$\sup_{\|r\|_{\nu^+,\pi}\le 1} \|A_r^{\pi'} - \widehat{A}_r\|_{d^\pi}^2 \le \frac{(4-3\gamma)L_{\max}^2}{2(1-\gamma)} \mathcal{D}_{\mathrm{KL}}(\pi\|\pi').$$

□

D.8 Reward-independent policy gradient approximation error bound

In this section, we show that SUAF can be reused within the specified trust region, with reward-uniform bounds on advantage and policy-gradient approximation errors.

Lemma D.12 (Reward-independent policy gradient approximation error bound)**.** *Given the reference policy π_θ in the trust region of π_{ref}, i.e., $\mathcal{D}(\pi_\theta\|\pi_{ref}) < \delta$, the approximation error of using SUAF of π_{ref} for policy gradient satisfies that*

$$\sup_{\|r\|_{\nu^+,\pi_{ref}}\le 1} \left\|\nabla_\theta J_r(\pi_\theta) - \widehat{G}_r\right\|_{F_\theta^\dagger}^2 < \frac{\gamma L_{\max}^2}{2(1-\gamma)}\delta,$$

where $\widehat{G}_r = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}}[\nabla_\theta \log \pi_\theta \widehat{A}_r]$ with $\widehat{A}_r = C_{\pi_\theta} A_r^{\pi_{ref}}$ is the approximated policy gradient and $F_\theta = \mathbb{E}_{d^{\pi_\theta}}[\nabla_\theta \log \pi_\theta (\nabla_\theta \log \pi_\theta)^T]$ denotes the Fisher matrix according to policy π_θ with $\dagger$ denoting pseudo-inverse.

Proof. Fix a reward r with $\|r\|_\nu \le 1$, and set

$$e_r := A_r^{\pi_\theta} - \widehat{A}_r, G_r(\theta) = \nabla_\theta J_r(\pi_\theta), b_r := G_r(\theta) - \widehat{G}_r(\theta) = \mathbb{E}_{d^{\pi_\theta}}[\nabla_\theta \log \pi_\theta e_r].$$

Write

$$\varepsilon_\delta := \left(\frac{\gamma L_{\max}^2}{2(1-\gamma)}\delta\right)^{1/2}. \tag{73}$$

By Theorem D.11, the advantage transfer bound yields

$$\|e_r\|_{d^{\pi_\theta}}^2 \le \frac{\gamma L_{\max}^2}{2(1-\gamma)} \mathcal{D}_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}) < \varepsilon_\delta^2. \tag{74}$$

For any vector u, Cauchy–Schwarz gives

$$|u^\top b_r| \le \left(\mathbb{E}_{d^{\pi_\theta}}[(u^\top \nabla_\theta \log \pi_\theta)^2]\right)^{1/2} \|e_r\|_{d^{\pi_\theta}} = \sqrt{u^\top F_\theta u}\, \|e_r\|_{d^{\pi_\theta}}. \tag{75}$$

Taking $u = F_\theta^\dagger b_r$ and using $F_\theta^\dagger F_\theta F_\theta^\dagger = F_\theta^\dagger$, we obtain $\|b_r\|_{F_\theta^\dagger}^2 \le \|b_r\|_{F_\theta^\dagger} \|e_r\|_{d^{\pi_\theta}}$, which further gives

$$\|b_r\|_{F_\theta^\dagger}^2 \le \|e_r\|_{d^{\pi_\theta}}^2 < \varepsilon_\delta^2.$$

Note that the above inequality holds for all reward function. Take the supremum and we prove the theorem. □

D.9 Proof of Corollary D.12

Corollary D.13 (Policy improvement using transferred SUAF)**.** *Under the conditions of Theorem D.12, fix a reward r with $\|r\|_\nu \le 1$ and define $u_r := F_\theta^\dagger \widehat{G}_r(\theta), s_r := \|\widehat{G}_r(\theta)\|_{F_\theta^\dagger}, \varepsilon_\delta := \left(\frac{\gamma L_{\max}^2}{2(1-\gamma)}\delta\right)^{1/2}$ Then $\nabla_\theta J_r(\pi_\theta)^\top u_r \ge s_r(s_r - \varepsilon_\delta)$. In particular, if $s_r > \varepsilon_\delta$, then u_r is a strict ascent direction, and there exists $\eta_0 > 0$ such that*

$$J_r(\pi_{\theta+\eta u_r}) > J_r(\pi_\theta), \qquad 0 < \eta < \eta_0. \tag{76}$$

Moreover, if $\theta \mapsto J_r(\pi_\theta)$ has an L-Lipschitz gradient on a convex neighborhood containing the update segment, strict improvement holds whenever

$$0 < \eta < \frac{2 s_r(s_r - \varepsilon_\delta)}{L \|u_r\|_2^2}. \tag{77}$$

Proof. Fix a reward r with $\|r\|_\nu \le 1$, by Theorem D.12, we have that

$$\|G_r - \widehat{G}_r\|_{F_\theta^\dagger} \le \varepsilon_\delta.$$

Therefore, Cauchy–Schwarz for the positive semi-definite form induced by $F_\theta^\dagger$ gives

$$\begin{aligned} G_r^\top u_r &= \widehat{G}_r^\top F_\theta^\dagger \widehat{G}_r + (G_r - \widehat{G}_r)^\top F_\theta^\dagger \widehat{G}_r \\ &\ge s_r^2 - \|G_r - \widehat{G}_r\|_{F_\theta^\dagger} s_r \\ &\ge s_r(s_r - \varepsilon_\delta). \end{aligned} \tag{78}$$

When $s_r > \varepsilon_\delta$, this directional derivative is strictly positive. Differentiability implies

$$J_r(\pi_{\theta+\eta u_r}) - J_r(\pi_\theta) = \eta G_r^\top u_r + o(\eta) \ge 0$$

for all sufficiently small $\eta > 0$. Under the additional smoothness assumption,

$$\begin{aligned} J_r(\pi_{\theta+\eta u_r}) - J_r(\pi_\theta) &\ge \eta G_r^\top u_r - \frac{L\eta^2}{2}\|u_r\|_2^2 \\ &\ge \eta s_r(s_r - \varepsilon_\delta) - \frac{L\eta^2}{2}\|u_r\|_2^2. \end{aligned} \tag{79}$$

The stated step-size condition makes this lower bound strictly positive. □

D.10 Proof of Corollary 4.3

Corollary D.14 (Policy-gradient error with truncated SUAF)**.** *Let $\pi = \pi_{\mathrm{ref}}$ and $\pi' = \pi_\theta$. Suppose $\mathcal{D}_{\mathrm{KL}}(\pi' \| \pi) < \delta$. Define the transferred rank-k approximation by*

$$\widehat{A}_r^k = C_{\pi'} r + \frac{\gamma}{1-\gamma} \sum_{i=1}^{k} \sigma_i^{+,\pi} (C_{\pi'} f_i^\pi) \langle g_i^\pi, r \rangle_{\nu^{+,\pi}},$$

and let $\widehat{G}_\theta^k(r) = \mathbb{E}_{d^{\pi'}}[\nabla_\theta \log \pi_\theta(a \mid s)\, \widehat{A}_r^k(s,a)]$. *For* $\eta_\delta = \min\{1, \sqrt{\delta/[2(1-\gamma)]}\}$, $B(\delta) = \min\{R_{\max}^2, 1 + R_{\max}^2 C_\gamma \sqrt{\delta}\}$ *with* $C_\gamma = \sqrt{(4+\gamma^{-1})/(2(1-\gamma)^2)}$,

$$\sup_{\|r\|_{\nu^{+,\pi'}} \le 1} \left\| \nabla_\theta \bar{J}_r(\pi_\theta) - \widehat{G}_\theta^k(r) \right\|_{F_\theta^\dagger}^2 \le B(\delta) \left[\sqrt{\frac{\gamma \delta L_{\max}^2}{2(1-\gamma)}} + \sqrt{\left(\frac{\gamma}{1-\gamma}\right)^2 \left(\sigma_{k+1}^{+,\pi}\right)^2 + L_{\max}^2 \eta_\delta} \right]^2 .$$

Proof. Write

$$c = \frac{\gamma}{1-\gamma}, \qquad \nu = \nu^{+,\pi}, \qquad \nu' = \nu^{+,\pi'},$$

and let

$$\mathcal{D} = \mathbb{E}_{s \sim d_S^{\pi'}} D_{\mathrm{KL}}\big(\pi'(\cdot \mid s) \| \pi(\cdot \mid s)\big) < \delta.$$

Step 1: Controlling the changes in sampling distributions. Let T_0 be independent of the trajectory with $\Pr(T_0 = t) = (1-\gamma)\gamma^t$ for $t \ge 0$. The distribution of (S_{T_0}, A_{T_0}) under a policy ρ is d^ρ. By the chain rule, the KL divergence between the laws of the trajectories stopped at T_0, under π' and π, equals

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi'} \left[D_{\mathrm{KL}}\big(\pi'(\cdot \mid S_t) \| \pi(\cdot \mid S_t)\big) \right] = \frac{\mathcal{D}}{1-\gamma}.$$

Pinsker's inequality therefore yield

$$\mathrm{TV}(d^{\pi'}, d^\pi) \le \min\left\{1, \sqrt{\frac{\mathcal{D}}{2(1-\gamma)}}\right\} \le \eta_\delta. \tag{80}$$

To control ν' and ν, independently draw $L \ge 1$ with $\Pr(L = \ell) = (1-\gamma)\gamma^{\ell-1}$, and set $N = T_0 + L$. The distribution of (S_N, A_N) under ρ is $\nu^{+,\rho} = d^\rho M^{+,\rho}$. Define

$$\epsilon_t = \mathbb{E}_{\pi'} \left[\mathrm{TV}\big(\pi'(\cdot \mid S_t), \pi(\cdot \mid S_t)\big) \right].$$

Jensen's and Pinsker's inequalities imply

$$\sum_{t=0}^{\infty} (1-\gamma)\gamma^t \epsilon_t^2 \le \frac{\mathcal{D}}{2}.$$

For each fixed horizon, replacing the action kernels one at a time bounds the total variation of the terminal distribution by the sum of the expected per-step action total variations under π'. Averaging over N consequently gives

$$\mathrm{TV}(\nu', \nu) \le \sum_{t=0}^{\infty} a_t \epsilon_t, \qquad a_t := \Pr(N \ge t).$$

Here

$$a_0 = 1, \qquad a_t = \gamma^t + t(1-\gamma)\gamma^{t-1}, \quad t \ge 1.$$

Weighted Cauchy–Schwarz yields

$$\mathrm{TV}(\nu', \nu)^2 \le \left(\sum_{t=0}^{\infty} \frac{a_t^2}{(1-\gamma)\gamma^t} \right) \left(\sum_{t=0}^{\infty} (1-\gamma)\gamma^t \epsilon_t^2 \right) \le \frac{4+\gamma^{-1}}{(1-\gamma)^2} \frac{\mathcal{D}}{2} \le C_\gamma^2 \delta.$$

Consequently, for every reward satisfying $\|r\|_{\nu'} \le 1$ and $\|r\|_\infty \le R_{\max}$,

$$\|r\|_\nu^2 \le \|r\|_{\nu'}^2 + R_{\max}^2\, \mathrm{TV}(\nu, \nu') \le 1 + R_{\max}^2 C_\gamma \sqrt{\delta}.$$

Since also $\|r\|_\nu^2 \le R_{\max}^2$, we obtain

$$\|r\|_\nu^2 \le B(\delta). \tag{81}$$

Step 2: Bounding the spectral residual on the reference reward ball. Temporarily fix a reward u with $\|u\|_\nu \le 1$. Define its omitted reward component by

$$u_{>k} = \sum_{i=k+1}^{R} \langle g_i^\pi, u\rangle_\nu g_i^\pi.$$

Orthonormality gives $\|u_{>k}\|_\nu \le 1$.

The state-action transition operator satisfies $P_\pi C_\pi = 0$, and hence $\mathcal{T}^{+,\pi} C_\pi = 0$. Moreover, $\nu(s,a) = \nu_S(s)\pi(a \mid s)$, so C_π is an orthogonal projection in $L^2(\nu)$. Every right singular function corresponding to a nonzero singular value therefore satisfies

$$C_\pi g_i^\pi = 0.$$

Indeed, for every $v \in L^2(\nu)$,

$$\langle g_i^\pi, C_\pi v\rangle_\nu = \frac{1}{\sigma_i^{+,\pi}} \langle f_i^\pi, \mathcal{T}^{+,\pi} C_\pi v\rangle_{d^\pi} = 0.$$

It follows that $C_\pi u_{>k} = 0$ and

$$e_{k,u} := A_u^\pi - A_{u,k}^\pi = c \sum_{i=k+1}^{R} \sigma_i^{+,\pi} f_i^\pi \langle g_i^\pi, u\rangle_\nu = A_{u_{>k}}^\pi. \tag{82}$$

Because $\|u_{>k}\|_\nu \le 1$, the definition of $L_{\max}$ bounds the action span of $e_{k,u}$ by $L_{\max}$. Its conditional action mean under π is zero. Therefore

$$|e_{k,u}(s,a)| \le L_{\max}.$$

The spectral approximation bound also gives

$$\|e_{k,u}\|_{d^\pi}^2 \le c^2 \big(\sigma_{k+1}^{+,\pi}\big)^2.$$

Using equation 80, we conclude

$$\begin{aligned} \|e_{k,u}\|_{d^{\pi'}}^2 &\le \|e_{k,u}\|_{d^\pi}^2 + L_{\max}^2 \,\mathrm{TV}(d^{\pi'}, d^\pi) \\ &\le c^2 \big(\sigma_{k+1}^{+,\pi}\big)^2 + L_{\max}^2 \eta_\delta. \end{aligned} \tag{83}$$

Step 3: Combining transfer and truncation errors. Let $\widehat{A}_u^\infty = C_{\pi'} Q_u^\pi$ denote the transferred full-rank representation. Since $C_{\pi'}$ annihilates state-dependent functions, the transferred truncated representation satisfies $\widehat{A}_u^k = C_{\pi'} A_{u,k}^\pi$. Thus

$$\widehat{A}_u^\infty - \widehat{A}_u^k = C_{\pi'} e_{k,u}.$$

The operator $C_{\pi'}$ is an orthogonal projection in $L^2(d^{\pi'})$, so

$$\|\widehat{A}_u^\infty - \widehat{A}_u^k\|_{d^{\pi'}} \le \|e_{k,u}\|_{d^{\pi'}}.$$

Combining equation 83 with the full-rank transfer theorem and the triangle inequality gives

$$\begin{aligned} \|A_u^{\pi'} - \widehat{A}_u^k\|_{d^{\pi'}} &\le \|A_u^{\pi'} - \widehat{A}_u^\infty\|_{d^{\pi'}} + \|\widehat{A}_u^\infty - \widehat{A}_u^k\|_{d^{\pi'}} \\ &\le L_{\max} \sqrt{\frac{\gamma\delta}{2(1-\gamma)}} + \sqrt{c^2 \big(\sigma_{k+1}^{+,\pi}\big)^2 + L_{\max}^2 \eta_\delta}. \end{aligned}$$

For $\bar{J}_u = (1-\gamma) J_u$, the policy-gradient identity and Cauchy–Schwarz imply

$$\Big\| \nabla_\theta \bar{J}_u(\pi_\theta) - \widehat{G}_\theta^k(u) \Big\|_{F_\theta^\dagger} \le \|A_u^{\pi'} - \widehat{A}_u^k\|_{d^{\pi'}}.$$

Indeed, writing $s_\theta = \nabla_\theta \log \pi_\theta(a \mid s)$, the gradient error is $\mathbb{E}_{d^{\pi'}}[s_\theta (A_u^{\pi'} - \widehat{A}_u^k)]$, and $F_\theta = \mathbb{E}_{d^{\pi'}}[s_\theta s_\theta^\top]$.

Step 4: Returning to the target reward class. Now let $\|r\|_{\nu'} \le 1$ and $\|r\|_\infty \le R_{\max}$. If $t := \|r\|_\nu > 0$, set $u = r/t$. Then $\|u\|_\nu = 1$. Both the exact policy gradient and its SUAF estimate are linear in the reward, so their error for r equals t times their error for u. Using equation 81, we obtain

$$\left\|\nabla_\theta \bar{J}_r(\pi_\theta) - \widehat{G}^k_\theta(r)\right\|^2_{F^\dagger_\theta} \le B(\delta)\left[L_{\max}\sqrt{\frac{\gamma\delta}{2(1-\gamma)}} + \sqrt{c^2\left(\sigma^{+,\pi}_{k+1}\right)^2 + L^2_{\max}\eta_\delta}\right]^2.$$

The case $t = 0$ is immediate under the full-support assumption. Taking the supremum over $\{r : \|r\|_{\nu'} \le 1,\ \|r\|_\infty \le R_{\max}\}$ completes the proof. □

E Algorithms

E.1 Algorithm to Estimate SUAF

We provide the derivations omitted from the main text.

Reverse conditional and marginal distributions. For one-step transitions $(s,a) \sim d^\pi$ and $(s',a') \sim P_\pi(\cdot|s,a) = \pi(a'|s')T(s'|s,a)$, the reverse conditional is

$$p_\pi(s,a|s',a') = \frac{d^\pi(s,a)P_\pi(s',a'|s,a)}{\nu^\pi_1(s',a')},$$

with marginal

$$\nu^\pi_1(\cdot,\cdot) := \sum_{s,a} d^\pi(s,a)P_\pi(\cdot,\cdot|s,a).$$

All density ratios reduce to one when d^π is stationary.

Adjoint interpretation. Denote by $(\mathcal{T}^{+,\pi})^*$ the adjoint of $\mathcal{T}^{+,\pi}$. Then $b = (\mathcal{T}^{+,\pi})^* f_\theta$ and $h - \mathbb{E}_\pi[h|s] = \mathcal{T}^{+,\pi} g_\omega$, which yields the two equivalent expressions in equation 11.

Backward prediction with marginal correction. When the sampling distribution is not stationary, we define

$$w_1(s',a') = \frac{\nu^\pi_1(s',a')}{d^\pi(s',a')}, \qquad w_+(s',a') = \frac{d^\pi(s',a')}{\nu^{+,\pi}(s',a')}.$$

Writing $b = w_+ v$, the backward recursion becomes

$$v(s',a') = w_1(s',a')\,\mathbb{E}[(1-\gamma)f_\theta(s,a) + \gamma v(s,a) \mid s',a'],$$

and $\widehat{b}_\theta = w_+ v_\eta$.

Surrogate backward loss. Below, we treat $\mathrm{sg}(\cdot)$ as a stop-gradient constant. There are only two key identities:

$$\mathbb{E}_\nu\left[\frac{d^\pi}{\nu}h\right] = \mathbb{E}_{d^\pi}[h], \qquad \mathbb{E}_{\nu_B}[w_1 h] = \mathbb{E}_P h,$$

where P is the joint distribution after one step starting from d^π.

Let

$$T(s,a) = (1-\gamma)f_\theta(s,a) + \gamma v_{\bar{\eta}}(s,a).$$

Taking one step from d^π yields

$$P(s',a') = \sum_{s,a} d^\pi(s,a)P_\pi(s',a' \mid s,a).$$

Define

$$\nu_B(s',a') = \gamma P(s',a') + (1-\gamma)d_0(s',a'), \qquad w_1(s',a') = \frac{P(s',a')}{\nu_B(s',a')}.$$

This is exactly the density ratio. The original backward loss with density ratio can be written in weighted least-squares form

$$\mathcal{L}_{\mathrm{B}}(\eta) = \frac{1}{2}\mathbb{E}_{(s',a')\sim\nu_B}\left[\|v_\eta(s',a') - w_1(s',a')\,\mathrm{sg}(\mathbb{E}_{P_\pi}[T(s,a) \mid s',a'])\|^2\right] + \text{const.}$$

where the constant is independent of v_η. Expanding the square:

$$\mathcal{L}_{\mathrm{B}} = \frac{1}{2}\mathbb{E}_{\nu_B}\|v_\eta\|^2 - \mathbb{E}_{\nu_B}\left[v_\eta^\top w_1 \operatorname{sg}(\mathbb{E}_{P_\pi}[T \mid s', a'])\right] + \text{const.}$$

Using $w_1\nu_B = P$, the second term becomes

$$\mathbb{E}_{\nu_B}\left[v_\eta^\top w_1 \operatorname{sg}(\mathbb{E}_{P_\pi}[T \mid s', a'])\right] = \mathbb{E}_{(s,a)\sim d^\pi,(s',a')\sim P_\pi}\left[v_\eta(s', a')^\top \operatorname{sg}(T(s, a))\right].$$

The first term splits as:

$$\mathbb{E}_{\nu_B}\|v_\eta\|^2 = \gamma\mathbb{E}_P\|v_\eta\|^2 + (1-\gamma)\mathbb{E}_{d_0}\|v_\eta\|^2.$$

Therefore

$$\mathcal{L}_{\mathrm{B}} = \frac{\gamma}{2}\mathbb{E}_P\|v_\eta\|^2 - \mathbb{E}_P\left[v_\eta^\top \operatorname{sg}(T)\right] + \frac{1-\gamma}{2}\mathbb{E}_{d_0}\|v_\eta\|^2 + \text{const.}$$

Now complete the square:

$$\frac{\gamma}{2}\mathbb{E}_P\left\|v_\eta - \frac{\operatorname{sg}(T)}{\gamma}\right\|^2 = \frac{\gamma}{2}\mathbb{E}_P\|v_\eta\|^2 - \mathbb{E}_P\left[v_\eta^\top \operatorname{sg}(T)\right] + \frac{1}{2\gamma}\mathbb{E}_P\|\operatorname{sg}(T)\|^2.$$

The last term is independent of v_η, hence

$$\mathcal{L}_{\mathrm{B}} = \frac{\gamma}{2}\mathbb{E}_{(s,a)\sim d^\pi,(s',a')\sim P_\pi}\left\|v_\eta(s', a') - \frac{\operatorname{sg}\big((1-\gamma)f_\theta(s, a) + \gamma v_{\bar\eta}(s, a)\big)}{\gamma}\right\|^2 + \frac{1-\gamma}{2}\mathbb{E}_{d_0}\|v_\eta\|^2 + \text{const.}$$

This is exactly $\widehat{\mathcal{L}}_{\mathrm{B}}(\eta)$, up to a constant independent of v_η.

Taking the stationary point with respect to $v_\eta(s', a')$:

$$\gamma P(s', a')\left(v_\eta - \frac{\mathbb{E}[T \mid s', a']}{\gamma}\right) + (1-\gamma)d_0(s', a')v_\eta = 0,$$

i.e.,

$$\big(\gamma P + (1-\gamma)d_0\big)v_\eta = P\,\mathbb{E}[T \mid s', a'].$$

Thus

$$v_\eta(s', a') = \frac{P(s', a')}{\gamma P(s', a') + (1-\gamma)d_0(s', a')}\mathbb{E}[T \mid s', a'] = w_1(s', a')\mathbb{E}[T \mid s', a'].$$

This is exactly the same fixed point as the original loss with density ratio. When d^π is stationary and $d_0 = d^\pi$, $P = d^\pi$, $\nu_B = d^\pi$, so $w_1 \equiv 1$.

Surrogate g-loss. Let $\nu = \nu^{+,\pi}$. The original g-loss with density ratio can be written as

$$\mathcal{L}_g(\omega) = \frac{1}{2}\mathbb{E}_\nu\left[g_\omega^\top \operatorname{sg}[C_f]g_\omega\right] - \mathbb{E}_\nu\left[\frac{d^\pi}{\nu}g_\omega^\top \operatorname{sg}(v_\eta)\right] + \text{const.}$$

The second term directly uses the density ratio identity:

$$\mathbb{E}_\nu\left[\frac{d^\pi}{\nu}g_\omega^\top \operatorname{sg}(v_\eta)\right] = \sum_{s,a}\nu(s, a)\frac{d^\pi(s, a)}{\nu(s, a)}g_\omega^\top \operatorname{sg}(v_\eta(s, a)) = \mathbb{E}_{d^\pi}\left[g_\omega^\top \operatorname{sg}(v_\eta)\right].$$

Therefore

$$\mathcal{L}_g(\omega) = \frac{1}{2}\mathbb{E}_{\nu^{+,\pi}}\left[g_\omega^\top \operatorname{sg}[C_f]g_\omega\right] - \mathbb{E}_{d^\pi}\left[g_\omega^\top \operatorname{sg}(v_\eta)\right] + \text{const.}$$

This is exactly

$$\widehat{\mathcal{L}}_g(\omega) = -\mathbb{E}_{(s,a)\sim d^\pi}\left[g_\omega^\top \operatorname{sg}(v_\eta(s, a))\right] + \frac{1}{2}\mathbb{E}_{\nu^{+,\pi}}\left[g_\omega^\top \operatorname{sg}[C_f]g_\omega\right].$$

up to a constant independent of g_ω.

Equivalently, if the original g-loss is written in weighted least-squares form

$$\mathcal{L}_g(\omega) = \frac{1}{2}\mathbb{E}_\nu\left\|g_\omega - \frac{d^\pi}{\nu}\operatorname{sg}(C_f^{-1}v_\eta)\right\|^2_{\operatorname{sg}[C_f]} + \text{const},$$

expanding also gives the same result: the quadratic term remains under ν, the density ratio in the cross term is cancelled by ν, becoming an expectation under d^π.

Differentiating with respect to g_ω, treating $\operatorname{sg}[C_f]$ as a constant matrix, yields

$$\mathbb{E}_{\nu^{+,\pi}}[\operatorname{sg}[C_f]g_\omega] = \mathbb{E}_{d^\pi}[\operatorname{sg}(v_\eta)].$$

This is the same stationary condition as the original loss with density ratio.

Main body of algorithm For simplicity, we denote $X = (S, A)$ and $Y_1 = (S', A')$. Based on the new notations, we describe the algorithm of estimating SUAF with neural networks and bidirectional TD as follows:

Algorithm 1 Neural Bidirectional TD Soft-HGR

Require: Fixed reference policy π; rank k; $0 < \gamma < 1$; outer iterations M; first-step replay $\mathcal{D}_{d_0}$, on-policy trajectory replay $\mathcal{D}_{\mathrm{tr}}$ and independent calibration/diagnostic replay $\mathcal{D}_{\mathrm{val}}$; TD steps H_F, H_B, feature steps H_f, H_g, learning rates, and target rates τ_F, τ_B.
Optimized networks: feature parameters θ, ω; forward predictor $h_\xi : \mathcal{X} \to \mathbb{R}^k$ and backward auxiliary predictor $v_\eta : \mathcal{X} \to \mathbb{R}^k$; Polyak targets $\bar{\xi}, \bar{\eta}$.
Derived objects: centered features f_θ, g_ω, sample moments, and final $k \times k$ whitening/SVD transforms; no feature tables.

Prepare one-step pairs (X, Y_1) and positive-lag pairs (X, Y^+) using the sampling rule below.
Initialize nonzero feature networks and predictors; set $\bar{\xi} \leftarrow \xi$, $\bar{\eta} \leftarrow \eta$.
for $m = 1, \dots, M$ **do**
  Freeze ω and its centering statistics.
  for $h = 1, \dots, H_F$ **do**
    Sample a minibatch (X, Y_1) from the one-step replay.
    $Z_F \leftarrow \mathrm{sg}[(1-\gamma) g_\omega(Y_1) + \gamma h_{\bar{\xi}}(Y_1)]$.
    $\xi \leftarrow \xi - \alpha_F \nabla_\xi \widehat{\mathbb{E}} \| h_\xi(X) - Z_F \|_2^2$.
    $\bar{\xi} \leftarrow (1-\tau_F)\bar{\xi} + \tau_F \xi$.
  end for
  Estimate $\widehat{C}_g \leftarrow \widehat{\mathbb{E}}_{Y^+}[g_\omega(Y^+) g_\omega(Y^+)^{\mathrm{T}}]$.
  for $j = 1, \dots, H_f$ **do**
    Sample $X \sim d^\pi$; evaluate the action-centered feature $f_\theta(X)$.
    $\mathcal{L}_f \leftarrow \widehat{\mathbb{E}}_X \left[-f_\theta(X)^{\mathrm{T}} \mathrm{sg}[h_\xi(X)] + \frac{1}{2} f_\theta(X)^{\mathrm{T}} \mathrm{sg}[\widehat{C}_g] f_\theta(X) \right]$.
    $\theta \leftarrow \theta - \alpha_f \nabla_\theta \mathcal{L}_f$. ▷ Only the left feature network is updated
  end for
  Freeze the updated θ.
  for $h = 1, \dots, H_B$ **do**
    Sample a minibatch (X, Y_1) from the same one-step replay.
    $Z_B \leftarrow \mathrm{sg}\big((1-\gamma) f_\theta(X) + \gamma v_{\bar{\eta}}(X)\big)$
    $\widehat{\mathcal{L}}_{\mathrm{B}}(\eta) \leftarrow \frac{\gamma}{2} \mathbb{E} \left\| v_\eta(Y_1) - \frac{Z_B}{\gamma} \right\|^2 + \frac{1-\gamma}{2} \mathbb{E}_{\mathcal{D}_{d_0}} \| v_\eta \|^2$
    $\eta \leftarrow \eta - \alpha_B \nabla_\eta \widehat{\mathcal{L}}_{\mathrm{B}}$.
    $\bar{\eta} \leftarrow (1-\tau_B)\bar{\eta} + \tau_B \eta$.
  end for
  Estimate $\widehat{C}_f \leftarrow \widehat{\mathbb{E}}_X[f_\theta(X) f_\theta(X)^{\mathrm{T}}]$.
  for $j = 1, \dots, H_g$ **do**
    Sample $Y^+ \sim \nu^{+,\pi}$; evaluate the marginal-centered feature $g_\omega(Y^+)$.
    $\widehat{\mathcal{L}}_g \leftarrow -\widehat{\mathbb{E}}_X \left[g_\omega^{\mathrm{T}} \mathrm{sg}(v_\eta(X)) \right] + \frac{1}{2} \widehat{\mathbb{E}}_{Y^+} \left[g_\omega^{\mathrm{T}} \mathrm{sg}[C_f] g_\omega \right]$.
    $\omega \leftarrow \omega - \alpha_g \nabla_\omega \widehat{\mathcal{L}}_g$. ▷ Only the right feature network is updated
  end for
end for
Freeze all networks; estimate calibration moments using $\mathcal{D}_{\mathrm{val}}$.
Obtain $(\widehat{f}, \widehat{g}, \widehat{\Sigma})$ by the small-matrix canonicalization below.
return Canonical feature functions $\widehat{f}, \widehat{g}$, estimated singular values $\widehat{\Sigma}$, and predictors h_ξ, v_η.

Sampling distributions. Write $X_t = (S_t, A_t)$ and $P_\pi((s', a') \mid (s, a)) = T(s' \mid s, a)\pi(a' \mid s')$. All trajectories start from $S_0 \sim \mu$ and follow the fixed policy π. Assume the conditional expectations and second moments below are finite. Draw independent times T_0 and L with

$$\Pr(T_0 = t) = (1-\gamma)\gamma^t \quad (t \geq 0), \qquad \Pr(L = \ell) = (1-\gamma)\gamma^{\ell-1} \quad (\ell \geq 1).$$

Set $X = X_{T_0}$, $Y_1 = X_{T_0+1}$, and $Y^+ = X_{T_0+L}$. Then

$$X \sim d^\pi, \quad Y_1 \sim \nu_1^\pi = d^\pi P_\pi, \quad Y^+ \mid X \sim M^{+,\pi}(\cdot \mid X), \quad Y^+ \sim \nu^{+,\pi} = d^\pi M^{+,\pi}.$$

Replay sampling must implement these weights. Uniformly sampling all stored transitions does not generally do so. Finite trajectories require an explicit tail treatment. Rejecting unavailable lags or

clipping them changes the target distribution. For episodic tasks, an absorbing-state convention must agree with the MDP definition. TD uses one-step pairs for the predictors. Positive-lag samples are still used for right-marginal moments, density-ratio estimation, and calibration.

Centering and the soft-HGR objective. Let $\widetilde{f}_\theta, \widetilde{g}_\omega : \mathcal{X} \to \mathbb{R}^k$ be raw network outputs. The feature functions used throughout the algorithm are

$$f_\theta(s,a) = \widetilde{f}_\theta(s,a) - \mathbb{E}_{\widetilde{a}\sim\pi(\cdot|s)}[\widetilde{f}_\theta(s,\widetilde{a})],$$
$$g_\omega(y) = \widetilde{g}_\omega(y) - \mathbb{E}_{Y^+\sim\nu^{+,\pi}}[\widetilde{g}_\omega(Y^+)].$$

Action integrals are approximated with policy samples and marginal means with reference samples. During a feature update, gradients pass through the centering operation of that feature. Predictor targets and the opposite covariance are detached. The population objective is

$$\mathcal{J}(\theta,\omega) = \mathbb{E}_{X,Y^+}[f_\theta(X)^{\mathrm{T}} g_\omega(Y^+)] - \tfrac{1}{2}\operatorname{tr}(C_f C_g), \quad C_f = \mathbb{E}_{d^\pi}[f_\theta f_\theta^{\mathrm{T}}], \quad C_g = \mathbb{E}_{\nu^{+,\pi}}[g_\omega g_\omega^{\mathrm{T}}].$$

Conditional centering restricts the learned dependence to the additional effect of the action given the state. Empirical centering and covariance estimates introduce sampling error. In particular, a noisy action-average estimate also perturbs second moments. Increasing the action-sample budget or using independent replicas for moment estimation can address this issue.

Why the two TD predictors provide the feature gradients. For fixed feature networks, the intended predictor solutions are

$$h(x) = \int M^{+,\pi}(y \mid x) g_\omega(y)\, dy, \qquad v(y) = \frac{\int d^\pi(x) M^{+,\pi}(y \mid x) f_\theta(x)\, dx}{d^\pi(y)}.$$

Integrals denote sums in the tabular case. The positive-lag resolvent identity gives

$$h(x) = \mathbb{E}[(1-\gamma) g_\omega(Y_1) + \gamma h(Y_1) \mid X = x],$$
$$v(y) = w_1(y)\mathbb{E}[(1-\gamma) f_\theta(X) + \gamma v(X) \mid Y_1 = y].$$

Both conditional expectations here use the one-step joint law $d^\pi(x)P_\pi(y \mid x)$. The output correction, in contrast, yields the positive-lag backward conditional expectation:

$$w_+(y)v(y) = \mathbb{E}[f_\theta(X) \mid Y^+ = y].$$

Consequently,

$$\mathbb{E}_{X,Y^+}[f_\theta(X)^{\mathrm{T}} g_\omega(Y^+)] = \mathbb{E}_{d^\pi}[f_\theta^{\mathrm{T}} h] = \mathbb{E}_{\nu^{+,\pi}}[g_\omega^{\mathrm{T}} w_+ v].$$

With exact predictors, ratios, centering, and moments, $\mathcal{L}_f$ and $\mathcal{L}_g$ therefore yield the negative block gradients of $\mathcal{J}$. While updating f_θ, h depends only on the fixed g_ω. While updating g_ω, v depends only on the fixed f_θ. Neither update requires differentiating through a TD solution. Finite TD fitting and statistical errors produce approximate block gradients. The construction alone does not establish convergence of nonlinear TD or global optimization of the neural objective.

Discounted-return post-calibration. After representation learning, we freeze the feature networks f_θ and g_ω and recalibrate their low-dimensional operator using an independent set of trajectory snippets. For each anchor $X_n = Z_{t_n}$, define the truncated discounted feature return

$$Y^g_{n,H} = \sum_{\ell=1}^{H} (1-\gamma)\gamma^{\ell-1} \bar{g}_\omega(Z_{t_n+\ell}), \tag{84}$$

where $\bar{g}_\omega$ is centered under the empirical positive-lag marginal. Equivalently, the trajectory snippets define the Rao–Blackwellized positive-lag joint estimator

$$\widehat{J}_H(x,y) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{X_n = x\} \sum_{\ell=1}^{H} (1-\gamma)\gamma^{\ell-1} \mathbf{1}\{Z_{t_n+\ell} = y\}. \tag{85}$$

We normalize $\widehat{J}_H$ by $\sum_{\ell=1}^{H}(1-\gamma)\gamma^{\ell-1} = 1-\gamma^H$. The omitted tail mass is therefore γ^H.

Let $\widehat{d}$ and $\widehat{\nu}$ denote the two marginals of $\widehat{J}_H$. Using the action-centered left feature $\bar{f}_\theta$ and marginal-centered right feature $\bar{g}_\omega$, we estimate

$$\widehat{C}_f = \mathbb{E}_{\widehat{d}}[\bar{f}_\theta(X)\bar{f}_\theta(X)^\top], \tag{86}$$

$$\widehat{C}_g = \mathbb{E}_{\widehat{\nu}}[\bar{g}_\omega(Y)\bar{g}_\omega(Y)^\top], \tag{87}$$

$$\widehat{K}_H = \mathbb{E}_{\widehat{J}_H}[\bar{f}_\theta(X)\bar{g}_\omega(Y)^\top] = \frac{1}{N}\sum_{n=1}^{N} \bar{f}_\theta(X_n)(Y^g_{n,H})^\top. \tag{88}$$

Unlike a single geometric-lag endpoint, the discounted feature return integrates over all future lags in the snippet and removes the additional geometric-lag sampling variance.

We use ridge-regularized whitening maps

$$W_f^{\lambda_f} = (\widehat{C}_f + \lambda_f I)^{-1/2}, \qquad W_g^{\lambda_g} = (\widehat{C}_g + \lambda_g I)^{-1/2}, \tag{89}$$

and compute the thin SVD

$$W_f^{\lambda_f}\widehat{K}_H W_g^{\lambda_g} = U\widehat{\Sigma}V^\top. \tag{90}$$

The corresponding calibrated features are

$$\widehat{f}(x) = U^\top W_f^{\lambda_f}\bar{f}_\theta(x), \qquad \widehat{g}(y) = V^\top W_g^{\lambda_g}\bar{g}_\omega(y). \tag{91}$$

Because of ridge regularization, these features satisfy

$$W_f^{\lambda_f}(\widehat{C}_f + \lambda_f I)(W_f^{\lambda_f})^\top = I, \tag{92}$$

and analogously for g. They are not exactly whitened with respect to $\widehat{C}_f$ and $\widehat{C}_g$ when the corresponding ridge parameters are nonzero.

Optionally, we continuously shrink the calibrated singular values:

$$\widetilde{\sigma}_i = a_i\widehat{\sigma}_i, \qquad a_i = \frac{\widehat{\sigma}_i^2}{\widehat{\sigma}_i^2 + \tau^2}. \tag{93}$$

The resulting rank-k operator is

$$\widehat{\mathcal{T}}^{+,k} r = \sum_{i=1}^{k} \widetilde{\sigma}_i\, \widehat{f}_i\, \langle \widehat{g}_i, r\rangle_{\widehat{\nu}}\,. \tag{94}$$

The ridge parameters (λ_f, λ_g) are selected on an independent calibration-validation split. After fixing them, the shrinkage parameter τ is selected on the same validation criterion without using any downstream task reward. The reported advantage errors are evaluated only after all calibration hyperparameters have been fixed.

What is learned. The returned functions approximate the reward-independent positive-lag, action-centered operator through

$$(\mathcal{T}^{+,\pi} r)(s,a) \approx \widehat{f}(s,a)^{\mathrm{T}}\widehat{\Sigma}\mathbb{E}_{Y^+\sim\nu^{+,\pi}}[\widehat{g}(Y^+)r(Y^+)].$$

For a downstream reward, the corresponding advantage estimate is

$$\widehat{A}^\pi_r(s,a) = C_\pi r(s,a) + \frac{\gamma}{1-\gamma}\widehat{f}(s,a)^{\mathrm{T}}\widehat{\Sigma}\mathbb{E}_{\nu^{+,\pi}}[\widehat{g}\, r], \qquad C_\pi r(s,a) = r(s,a) - \mathbb{E}_{a'\sim\pi(\cdot|s)}[r(s,a')].$$

Training uses no downstream reward. The two predictor networks propagate feature information through one-step transitions. Soft-HGR updates the feature networks and canonicalization fixes their final coordinates. Sample-efficiency gains over direct positive-lag training require empirical validation.

F Environment and Evaluation Details

This section specifies the finite-state environments used for the spectral diagnostics and one-step reward-transfer experiments. In every experiment, the transition kernel, initial distribution, and reference policy are fixed before the reward is selected. Consequently, changing a downstream reward does not change the data-generating process or the learned representation. Unless stated otherwise, we use $\gamma = 0.9$.

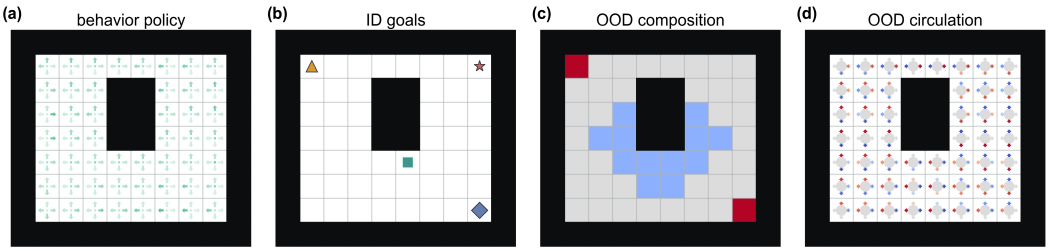


Figure 23: Behavior policy, environment, in-distribution and out-of-distribution tasks for grid world.

F.1 Discrete Finite Environments

F.1.1 Navigation environments

Both navigation domains have the five actions $\{\text{UP}, \text{RIGHT}, \text{DOWN}, \text{LEFT}, \text{STAY}\}$. Transitions are deterministic. A movement into a wall or outside the map leaves the agent in its current cell. The STAY action also leaves the state unchanged. The initial distribution μ is uniform over all free cells. Rewards do not terminate an episode: we evaluate the corresponding continuing discounted MDP exactly.

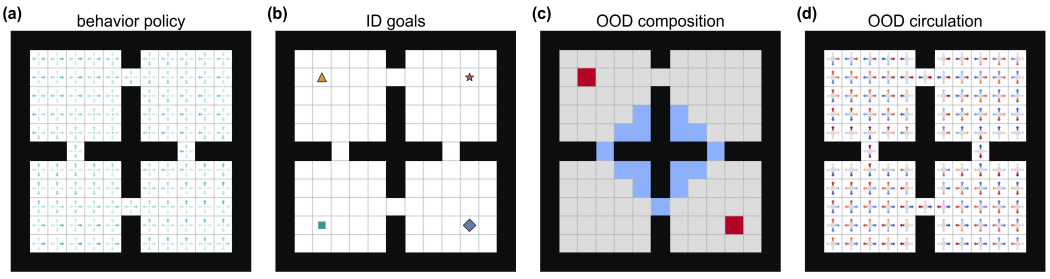


Figure 24: Behavior policy, environment, in-distribution and out-of-distribution tasks for Four-room.

Grid World. As is shown in Figure 23, Grid World uses a 7×8 canvas. The six cells

$$\{(i, j) : i \in \{1, 2, 3\},\ j \in \{3, 4\}\}$$

are blocked, leaving 50 states and 250 state–action pairs. Coordinates are zero-indexed from the top-left corner. The four in-distribution (ID) goal locations are top-left $(0, 0)$, top-right $(0, 7)$, center $(4, 4)$, and bottom-right $(6, 7)$. The population positive-lag DCDM has nonzero rank 49 under the reference policy used below.

Four-room. As is shown in Figure 24, Four-room uses an 11×11 canvas. An interior vertical wall lies at column 5 and an interior horizontal wall lies at row 5, with doorways at $(1, 5)$, $(8, 5)$, $(5, 2)$, and $(5, 8)$. Removing the wall cells leaves 104 states and 520 state–action pairs. The four ID goals are placed at $(1, 1)$, $(1, 9)$, $(9, 1)$, and $(9, 9)$. The corresponding population DCDM rank is 103.

F.1.2 Reference policies and discounted distributions

The navigation rank-control figures use one reproducible, full-support *fixed balanced* policy for each environment. Let π_{sm} be a smooth random policy obtained by applying a softmax to fixed-seed Fourier features of the cell coordinates. We use

$$\pi_0(a \mid s) = 0.2 \times \frac{1}{5} + 0.8\, \pi_{\text{sm}}(a \mid s).$$

Table 9: Finite-state environments used in the reward-transfer experiments. The DCDM rank is computed for the fixed reference policy at $\gamma = 0.9$.

Environment	Grid size	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	DCDM rank	Reference policy
Grid World	7×8	50	5	49	5 balanced
Four-room	11×11	104	5	103	5 balanced
Taxi	5×5	500	6	399	5 balanced

The uniform component guarantees $\pi_0(a \mid s) \geq 0.04$. The same realized policy is used to collect reward-free data, define the SUAF operator, and evaluate all rewards within an environment. It is not resampled per reward or per retained rank. Experiments explicitly described as the uniform reward-free baseline protocol instead use $\pi_0(a \mid s) = 1/5$. Those results are not mixed with the fixed-balanced rank-control curves.

For a state–action pair $X_t = (S_t, A_t)$, the normalized discounted occupancy and positive-lag marginal are

$$d^{\pi_0}(x) = (1-\gamma) \sum_{t \geq 0} \gamma^t \Pr_{\mu,\pi_0}(X_t = x), \tag{95}$$

$$\nu^{+,\pi_0}(y) = \sum_x d^{\pi_0}(x) M^{+,\pi_0}(y \mid x), \tag{96}$$

where $M^{+,\pi_0}(y \mid x) = (1-\gamma) \sum_{\ell \geq 1} \gamma^{\ell-1} \Pr_{\pi_0}(X_\ell = y \mid X_0 = x)$. All task rewards are normalized once using the reference-policy outcome marginal,

$$\widetilde{r} = \frac{r}{\|r\|_{\nu^{+,\pi_0}}}, \qquad \|r\|^2_{\nu^{+,\pi_0}} = \sum_x \nu^{+,\pi_0}(x) r(x)^2. \tag{97}$$

The normalization is fixed before policy improvement and is not recomputed under the improved policy.

F.1.3 ID AND OOD REWARD FAMILIES

Sparse ID goals. For a goal cell g, the unnormalized reward is state-only,

$$r_g(s,a) = \mathbf{1}\{s = g\}, \tag{98}$$

for every action. Grid World uses the four goal cells listed above. Four-room uses one goal inside each room. These sparse goals form the ID family used while selecting the representation protocol.

OOD state composition. This reward composes two diagonal ID goals and introduces a signed component that was absent from the ID family:

$$r_{\mathrm{comp}}(s,a) = r_{\mathrm{TL}}(s,a) + r_{\mathrm{BR}}(s,a) - \tfrac{1}{2}\mathbf{1}\{s \in \mathcal{H}\}. \tag{99}$$

Here $\mathcal{H}$ contains all valid cells whose Manhattan distance to the geometric map center is no larger than the fifth order statistic. Ties may therefore make $|\mathcal{H}| > 5$. Thus this task tests a new signed combination of spatial events, rather than another single-goal readout.

OOD state–action circulation. Let c be the geometric map center, $p_s \in \mathbb{R}^2$ the row–column coordinate of state s, and

$$t_s = \frac{(p_{s,2} - c_2, -(p_{s,1} - c_1))}{\max\{\|p_s - c\|_2, 1\}}$$

the clockwise tangent direction. The circulation reward is

$$r_{\mathrm{circ}}(s,a) = t_s^\top \left(\mathbb{E}[S' \mid s,a] - p_s\right). \tag{100}$$

Unlike all ID goals, this reward depends explicitly on the action and favors clockwise motion around the map center. The equation 99 and equation 100 are normalized using equation 97 before evaluation.

F.1.4 CONTINUING TAXI

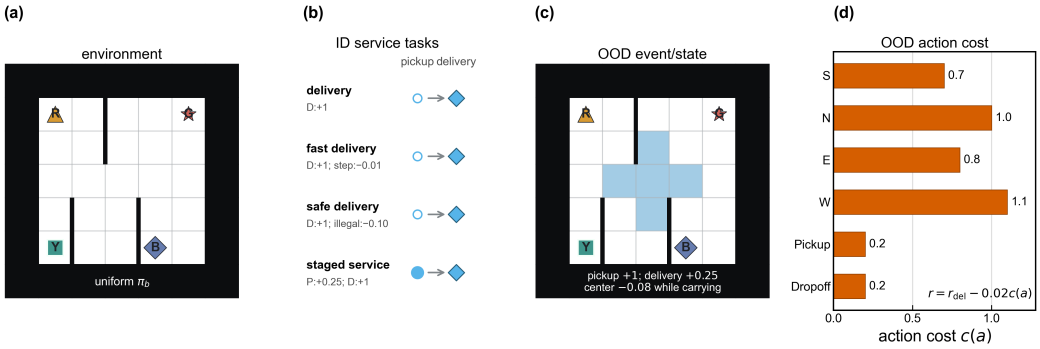


Figure 25: Environment, in-distribution and out-of-distribution tasks for taxi.

As is shown in Figure 25, Taxi uses the Taxi-v3 5×5 road map with 500 encoded states and six actions: four movements, pickup, and dropoff. A state records the taxi position, one of four passenger landmarks (or that the passenger is inside the taxi), and one of four destinations. The initial law is uniform over taxi positions and over passenger–destination pairs with distinct landmarks, giving support on $25 \times 4 \times 3 = 300$ encoded states. Successful delivery is followed by a draw from this same initial law, producing a continuing MDP without changing the Taxi-v3 within-task dynamics. The reference policy is uniform over the six actions, and the population DCDM rank is 399.

The four ID rewards are: successful delivery. Delivery with a 0.01 step cost. Delivery with a 0.10 illegal pickup/dropoff penalty, and delivery plus 0.25 for successful pickup. The three OOD rewards change the task structure: (i) pickup priority, 0.25 times delivery plus pickup; (ii) center hazard, delivery minus 0.08 whenever a passenger-carrying taxi is within Manhattan distance one of the map center; and (iii) action-dependent fuel cost, delivery minus 0.02 times the fixed action-cost vector $(0.7, 1.0, 0.8, 1.1, 0.2, 0.2)$. As in the navigation domains, every Taxi reward is normalized under ν^{+,π_0}, while dynamics, reset law, and reference policy remain unchanged.

Tasks We evaluate reward transfer while holding the MDP dynamics, state and action spaces, behavior policy, and offline replay distribution fixed. Thus, "OOD" refers exclusively to a shift in the reward family rather than to a dynamics or observation-distribution shift. In Grid World and Four-room, the ID tasks are four sparse single-goal rewards located in different regions of the environment. We construct two structurally distinct OOD tasks: a compositional reward that combines two distant goals with an unseen negative center hazard, and an action-dependent circulation reward that favors clockwise motion around the center of the map. In continuing Taxi, the ID tasks are four related service objectives: successful delivery, fast delivery with a step cost, safe delivery with an illegal-action penalty, and staged service with an intermediate pickup bonus. The Taxi OOD tasks alter the task structure more substantially by prioritizing pickup over delivery, penalizing carrying the passenger through the central region, or introducing action-dependent fuel costs. All rewards are normalized to unit ν^{+,π_b}-weighted norm, and the OOD rewards are excluded from model selection and hyperparameter tuning.

F.2 CONTINUOUS ENVIRONMENTS

We evaluate reward-independent representation learning and downstream reward transfer on continuous-control environments from ExORL and OGBench as shown in Figure 26. Within each domain, all downstream tasks share the same observation space, action space, transition dynamics, and offline transition dataset. The representation is learned without access to task rewards. At evaluation time, different control problems are created by changing only the downstream reward or goal. This protocol tests whether a single representation of the dynamics can be reused across multiple tasks.

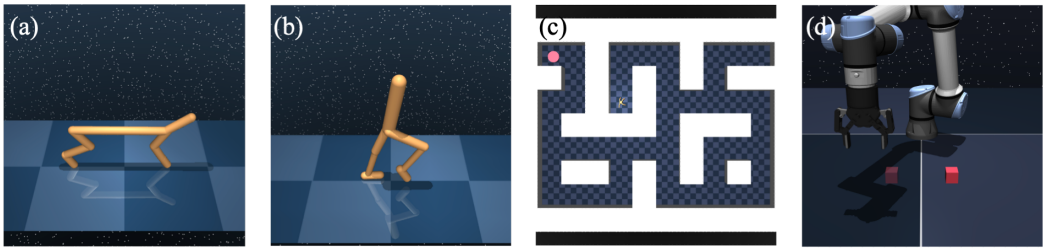


Figure 26: Environments of ExORL and OGBench.

ExORL locomotion environments. We use the *walker* and *cheetah* domains from ExORL. These environments evaluate transfer among locomotion objectives that share the same underlying physical system but favor different behaviors, velocities, or directions of motion.

For *walker*, we consider four tasks:

1. *flip*, which rewards completing a flipping motion;
2. *run*, which rewards fast forward locomotion;
3. *stand*, which rewards maintaining an upright stationary pose;
4. *walk*, which rewards forward locomotion at a walking speed.

The tasks therefore range from static stabilization to periodic locomotion and agile whole-body motion.

For *cheetah*, we consider four tasks:

1. *run*, which rewards fast forward locomotion;
2. *run backward*, which rewards fast locomotion in the reverse direction;
3. *walk*, which rewards slower forward locomotion;
4. *walk backward*, which rewards slower locomotion in the reverse direction.

The forward and backward tasks test whether the learned representation can be reused when the preferred direction of motion changes, while the walking and running tasks additionally vary the desired locomotion regime.

OGBench AntMaze Large Navigate. This domain requires a continuous-control agent to navigate between distant regions of a large maze. Each task is specified by a starting region and a target region. The maze geometry and transition dynamics remain fixed, but different start-goal pairs may require substantially different long-horizon routes. We consider five tasks:

1. bottom left $\rightarrow$ top right;
2. center $\rightarrow$ top left;
3. center $\rightarrow$ bottom right;
4. bottom right $\rightarrow$ center;
5. bottom left $\rightarrow$ center.

These tasks evaluate whether the learned representation captures the long-horizon connectivity of the maze well enough to support transfer among different navigation objectives.

OGBench Cube Single Play. This domain is a continuous robotic manipulation problem in which the agent must pick up a cube and place it at a specified target location. The robot, object, and contact dynamics are shared by all tasks; only the desired placement target changes. We evaluate five tasks:

Table 10: **Environment, replay, and representation dimensions.** A geometric positive lag has law $\Pr(L = \ell) = (1-\gamma)\gamma^{\ell-1}$ for $\ell \geq 1$. The Taxi model outputs 499 coordinates as an overparameterization diagnostic, while the population DCDM has rank 399. Coordinates beyond rank 399 do not correspond to additional nonzero population SUAF modes.

Hyperparameter	GridWorld	Four-room	Taxi
States / actions	$50/5$	$104/5$	$500/6$
State–action indices	250	520	3000
Population DCDM rank	49	103	399
Learned output dimension k	49	103	499
Discount γ	0.9	0.9	0.9
Reference policy	fixed balanced	fixed balanced	uniform
Initial law μ	uniform over free cells	uniform over free cells	Taxi-v3 law (300-state support)
Replay transitions	5×10^6	10^7	5×10^6
Trajectory horizon	100	100	100
Number of trajectories	50,000	100,000	50,000
Occupancy-anchor sampling	discounted	discounted	discounted
Positive-lag sampling	geometric$(1-\gamma)$	geometric$(1-\gamma)$	geometric$(1-\gamma)$
Minibatch size	1024	1024	1024

1. pick and place to the left;
2. pick and place to the front;
3. pick and place to the back;
4. pick and place along a diagonal direction;
5. pick and place along an off-diagonal direction.

Unlike the locomotion and navigation domains, successful transfer in this environment requires the representation to capture contact-rich manipulation dynamics, including both grasping and task-dependent object transport.

For every domain, *overall* denotes the arithmetic mean of the evaluation score over all downstream tasks in that domain.

G Experimental details

G.1 Data processing

Uncertainty estimation. For each experimental configuration, we repeat the experiment with N independent random seeds. Let x_i denote the evaluation metric from the i-th run at a given rank or training checkpoint. We compute the mean as $\bar{x} = N^{-1}\sum_{i=1}^{N} x_i$ and the sample standard deviation (std) as $s = \sqrt{(N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$. The 95% confidence interval (CI) for the mean is $\bar{x} \pm t_{0.975,N-1}s/\sqrt{N}$, where $t_{0.975,N-1}$ is the 97.5th percentile of Student's t-distribution with $N-1$ degrees of freedom. Standard deviations describe variability across runs, whereas confidence intervals quantify uncertainty in the estimated mean. Intervals are computed separately at each plotted point on the original metric scale, before any axis transformation, and represent pointwise rather than simultaneous confidence intervals.

G.2 Implementation details and hyperparameters of discrete environments

Tables 10–13 report the resolved settings used to produce the learned-representation and one-step-control results. We distinguish an *optimizer update* from an outer bidirectional iteration: one outer iteration performs, in order, one update of the forward predictor h, the left feature f, the reverse predictor v, and the right feature g, and therefore contains four optimizer updates. Unless noted otherwise, training tensors use single precision and the small-matrix calibration is performed in double precision.

Table 11: **Bidirectional-TD SUAF optimization.** The implementation uses trainable tables (embedding matrices), not hidden-layer MLPs. Thus the table dimensions below fully specify the representation architecture. Target copies of h and v have the same shapes but are not optimized directly.

Hyperparameter	GridWorld	Four-room	Taxi
Left feature $f(s,a)$	250×49	520×103	3000×499
Right feature $g(s)$	50×49	104×103	500×499
Forward predictor $h(s,a)$	250×49	520×103	3000×499
Reverse predictor $v(s,a)$	250×49	520×103	3000×499
Hidden layers / activation	none / none	none / none	none / none
Trainable scalar parameters	39,200	171,392	4,740,500
Initialization	$0.2\,\mathcal{N}(0, I)$ independently for f, g, h, v		
Feature ordering	sequential nested-SVD mask		
Action centering	exact expectation under the fixed reference policy		
Right-feature action dependence	tied across actions: $g(s,a) \equiv g(s)$		
Reverse-TD objective	boundary-corrected matched squared loss		
Reverse successor coefficient	$\gamma/2 = 0.45$		
Reverse boundary coefficient	$(1-\gamma)/2 = 0.05$		
Optimizer	four independent Adam optimizers, one for each of f, g, h, v		
Adam $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.999, 10^{-8})$		
Weight decay / AMSGrad	0 / false		
Initial learning rate	3×10^{-3} for all four optimizers		
Learning-rate schedule	multiply by 0.5 per 10^5 optimizer updates; floor 3×10^{-5}		
Gradient-norm clipping	10 (separately for each update)		
Polyak coefficient τ	0.2		
Optimizer updates / outer iteration	4 (h, f, v, then g)		
Training budget	630,000	2,000,000	100,000
Outer iterations	157,500	500,000	25,000
Diagnostic interval	10,000 updates	10,000 updates	10,000 updates

Table 12: **Reward-free checkpoint selection and post-hoc calibration.** For GridWorld and Four-room, $\lambda_f = s_f \operatorname{tr}(C_f)/k$ and $\lambda_g = s_g \operatorname{tr}(C_g)/k$. Shrinkage uses $\widetilde{\sigma}_i = \sigma_i^3/(\sigma_i^2 + \tau_\sigma^2)$ with $\tau_\sigma = s_\sigma \sigma_1$.

Hyperparameter	GridWorld	Four-room	Taxi
Calibration construction	discounted feature return	discounted feature return	sampled geometric positive lag
Calibration training pairs	500,000	500,000	500,000
Calibration validation pairs	250,000	250,000	500,000
Calibration rollout horizon	128	128	at most 100 within trajectory
Omitted geometric tail mass	1.39×10^{-6}	1.39×10^{-6}	–
Covariance inverse-sqrt tolerance	10^{-12}	10^{-12}	10^{-12}
Ridge-scale candidates	$\{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$		fixed 10^{-3}
Selected (s_f, s_g)	$(10^{-3}, 10^{-3})$	$(10^{-3}, 10^{-3})$	$(10^{-3}, 10^{-3})$
Shrinkage-scale candidates	$\{0, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$		not swept
Selected s_σ	0	0	0 (no shrinkage)
Small-matrix factorization	full SVD of the ridge-whitened $k \times k$ cross moment		
Selection metric	independent reward-free weighted operator discrepancy		
Selected optimizer updates	630k for every run	1860/1080/1490/1730/1310k	20/30/20/40k

The reverse predictor uses the same paired one-step transition (X, Y_1) in its quadratic and cross terms. Its boundary minibatch is sampled independently from the known $d_0(s,a) = \mu(s)\pi_0(a \mid s)$. No population transition matrix or population SUAF mode is supplied to training. Target tables are updated as $\bar{h} \leftarrow (1-\tau)\bar{h} + \tau h$ and analogously for v. The learning rate at optimizer update t is $\max\{3 \times 10^{-5}, 3 \times 10^{-3}\, 0.5^{t/10^5}\}$.

The calibration split is never used to optimize the raw feature tables. GridWorld and Four-room share the same independently generated calibration and validation joint distributions across representation seeds. Taxi first draws 10^6 independent calibration pairs from each seed's replay and uses a fixed half/half split. Taxi's ridge scale is fixed rather than selected by the two-stage navigation-domain validation sweep. This difference is reported explicitly rather than treating the two procedures as identical.

Table 13: **Downstream one-step control and shared settings of the learned baseline adapter.** "Exact" refers only to finite-MDP evaluation after reward-free representation training. It does not supply oracle quantities to SUAF training.

Hyperparameter	GridWorld	Four-room	Taxi
SUAF rank grid	$0, 5, \ldots, 45, 49$	$0, 10, \ldots, 100, 103$	$0, 20, \ldots, 380, 399, 400, 420, \ldots, 480, 499$
Boltzmann inverse temperatures	$\{4, 32, 256\}$		
Additional policy	greedy $\arg\max_a \widehat{A}(s, a)$		
Number of policy updates	1 No actor training and no feature refresh		
Reward normalization	$\|r\|_{\nu^+,\pi_0} = 1$, fixed under the reference policy		
Return computation	exact finite Bellman solve from μ		
Optimal-return denominator	exact value iteration, tolerance 10^{-12}		
Baseline pretraining updates	200,000	200,000	100,000
Baseline minibatch size	1024	1024	1024
Baseline hidden widths	$(256, 256)$	$(256, 256)$	$(256, 256)$
FB / One-Step FB latent dimension	49	103	499
Other-baseline latent dimension	50	104	500
Baseline reward-readout samples	120,000	120,000	population readout
Baseline control candidates	greedy; $\eta \in \{0.25, 1, 4, 16, 64, 256\}$; matched KL $\in \{0.01, 0.05, 0.1\}$		
Baseline evaluation interval	10,000	10,000	5,000 updates

Table 14: **Active method-specific baseline settings after applying the common discrete adapter.** All methods use a two-hidden-layer MLP with widths $(256, 256)$ and GELU activations, Optax Adam with learning rate 10^{-4}, target-update coefficient 0.005, normalized latents, and no image encoder. Goal mixtures are ordered as current / geometric same-trajectory / random. Here d denotes the environment-specific latent dimension in Table 13.

Method	Layer normalization	Goal mixture	Aggregation	Active extra setting
Laplacian	none	$0.2/0.5/0.3$	Q: mean	orthonormality coefficient 0
BYOL-γ	none	$0/1/0$	Q: mean	orthonormality coefficient 0
TD-JEPA	representation and predictor	$0.2/0.5/0.3$	representation: mean; Q: min	orthonormality 1; ϕ dimension $2d$
ICVF	none	$0/0.625/0.375$	Q: mean	expectile 0.5
HILP	none	$0/0.625/0.375$	representation: mean; Q: min	expectile 0.5; single-state latent reward
FB	none	$0.2/0.5/0.3$	representation: mean	orthonormality 0; reward temperature 10
One-Step FB	forward and backward representations	$0.2/0.5/0.3$	representation: mean; Q: min	orthonormality 0; reward temperature 10

For Laplacian, BYOL-γ, TD-JEPA, ICVF, HILP, FB, and One-Step FB, we start from the upstream configurations in the public One-Step-FB implementation (revision `fea347e`), replace every representation, predictor, value, and actor hidden-width tuple exposed by the configuration with $(256, 256)$, set the latent dimension as in Table 13, set $\gamma = 0.9$, and disable the continuous-action tanh squash. Our adapter never optimizes or evaluates the continuous actor: every action maximization is an exact enumeration of the finite action set. Actor-only coefficients are therefore inactive and omitted. The remaining active method-specific settings are listed in Table 14.

G.3 Continuous-Control Environments and Implementation Details

This section specifies the implementation used for the state-based ExORL domains Walker and Cheetah, and the state-based OGBench domains AntMaze Large Navigate and Cube Single Play. We align the data protocol, network widths, actor parameterization, optimization budget, zero-shot evaluation, and reporting convention with the public One-Step FB implementation. The deliberate SUAF-specific differences are the positive-lag operator, action-centering of the left feature, nested mode ordering, bidirectional TD predictors, and post-hoc canonicalization.

G.3.1 Reproducibility convention.

The resolved settings below follow the executable commands and defaults in the public One-Step FB repository rather than silently combining them with prose defaults. In particular, the released ExORL command uses a three-layer backward MLP of width 256 and the released agent enables LayerNorm in the forward and backward representations. Table 3 of the accompanying One-Step FB paper instead lists a two-layer backward MLP of width 512 and states that LayerNorm is disabled.

We use the released command/code configuration and record this discrepancy here. The upstream source and revision used for the final runs should be archived with the experiment artifacts.

G.3.2 Datasets and evaluation protocol

Table 15: **Dataset and evaluation protocol.** These values reproduce the state-based protocol of One-Step FB. "Final score" denotes averaging the evaluations at the last three checkpoints, not selecting the best checkpoint.

Hyperparameter	Walker	Cheetah	AntMaze Large	Cube Single Play
Benchmark	ExORL	ExORL	OGBench	OGBench
Offline behavior data	RND	RND	play	play
Representation-training transitions	5×10^6	5×10^6	10^6	10^6
Zero-shot inference transitions	10^5	10^5	10^5	10^5
Reward used for zero-shot inference	dense task reward	dense task reward	semi-sparse	semi-sparse
Discount γ	0.98	0.98	0.99	0.99
Maximum evaluation horizon	1000	1000	1000	200
Gradient steps	10^6	10^6	10^6	10^6
Evaluation interval	10^5	10^5	10^5	10^5
Episodes per task/checkpoint	50	50	50	50
Evaluation metric	undiscounted return	undiscounted return	success rate	success rate
Final reported checkpoints	800/900/1000k	800/900/1000k	800/900/1000k	800/900/1000k
Independent random seeds	8	8	8	8
Seed identifiers	stored verbatim in every run manifest			
Uncertainty summary	mean $\pm$ std.	mean $\pm$ std.	mean $\pm$ std.	mean $\pm$ std.

The ExORL training and zero-shot splits are generated from disjoint portions of the same exploratory dataset. OGBench uses its official independent train and validation files. Downstream rewards are never used to optimize the reward-free representation. They are first introduced when computing the task coordinate and evaluating the resulting latent-conditioned actor.

G.3.3 Network architecture

Let d denote the learned representation dimension. Our left SUAF feature $f(s, a)$ and forward TD predictor $h(s, a)$ use the upstream forward-network architecture. The right outcome feature $g(s)$ and reverse TD predictor $v(s, a)$ use the upstream backward-network architecture. The actor consumes the concatenation of the observation and reward coordinate $\beta \in \mathbb{R}^d$. Two independently initialized left heads are used when computing the actor score. The right representation is shared.

Table 16: **Resolved network architecture.** A tuple lists hidden-layer widths. The output layer is not included. The twin count applies to the left feature and its target copy.

Hyperparameter	Walker	Cheetah	AntMaze Large	Cube Single Play
Representation dimension d	50	50	128	128
Left feature f / predictor h	$(1024, 1024, 1024)$	$(1024, 1024, 1024)$	$(512, 512, 512, 512)$	$(512, 512, 512, 512)$
Right feature g / predictor v	$(256, 256, 256)$	$(256, 256, 256)$	$(512, 512, 512, 512)$	$(512, 512, 512, 512)$
Actor MLP	$(1024, 1024, 1024)$	$(1024, 1024, 1024)$	$(512, 512, 512, 512)$	$(512, 512, 512, 512)$
Activation	ReLU	ReLU	GELU	GELU
Twin left heads	2	2	2	2
Shared right head	yes	yes	yes	yes
LayerNorm in f, g, h, v	yes	yes	yes	yes
LayerNorm in actor	no	no	no	no
State-dependent actor std.	no	no	no	no
Constant actor log-std.	0	0	0	0
Actor final-layer init. scale	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Actor tanh transformation	no	no	yes	yes
Action clipping	environment bounds	environment bounds	$[-1, 1]$	$[-1, 1]$
Image encoder	none	none	none	none

G.3.4 EMPIRICAL ACTION-CENTERING

The offline datasets do not provide a callable behavior policy. We therefore construct a nonparametric empirical conditional action distribution once from the reward-free training split. We draw a fixed reference pool of $N_{\mathrm{ref}} = 32{,}768$ factual state–action rows. States are standardized coordinate-wise using training-set statistics. For a query state s, let $\mathcal{N}_K(s)$ contain the $K = 16$ reference states with the smallest squared Euclidean distance, and let a_j be the factual action attached to reference state s_j. We define $\widehat{\pi}_b(\mathrm{d}a \mid s) = \frac{1}{K} \sum_{j \in \mathcal{N}_K(s)} \delta_{a_j}(\mathrm{d}a)$, with $K = 16$ and action-center every raw left feature as

$$f_{\mathrm{c}}(s, a) = f_{\mathrm{raw}}(s, a) - \frac{1}{K} \sum_{j \in \mathcal{N}_K(s)} f_{\mathrm{raw}}(s, a_j). \tag{101}$$

For a training query originating from episode e, reference rows from the same original episode are masked before finding nearest neighbors. This cross-fitting rule prevents the factual action from being deterministically included in its own baseline and avoids the finite-K shrinkage caused by reusing the factual action. The neighbor search and the feature baseline are recomputed for the current query state, but the reference pool and its standardization statistics remain fixed throughout training. No behavior policy MLP is fitted.

Action-centering in equation 101 is used in the feature loss, reverse-TD target, calibration covariance, whitening/SVD, and all reported advantage errors. During actor differentiation the subtracted baseline is constant with respect to the proposed action at fixed s. Consequently it may be cached without changing the actor gradient. We still include it when logging numerical advantages, so the reported score is centered and not an arbitrary state-dependent offset.

Table 17: Hyperparameters of empirical action-centering, shared across the four continuous-control domains.

Hyperparameter	Value
Behavior model	nonparametric empirical kNN
Reference rows N_{ref}	32,768
Neighbors / action samples K	16
State metric	squared Euclidean after coordinate standardization
Neighbor weighting	uniform, $1/K$
Reference-pool sampling	uniform without replacement, fixed per seed
Same-episode reference rows	excluded during representation training
Factual action forcibly included	no
Gradient through selected baseline actions	no

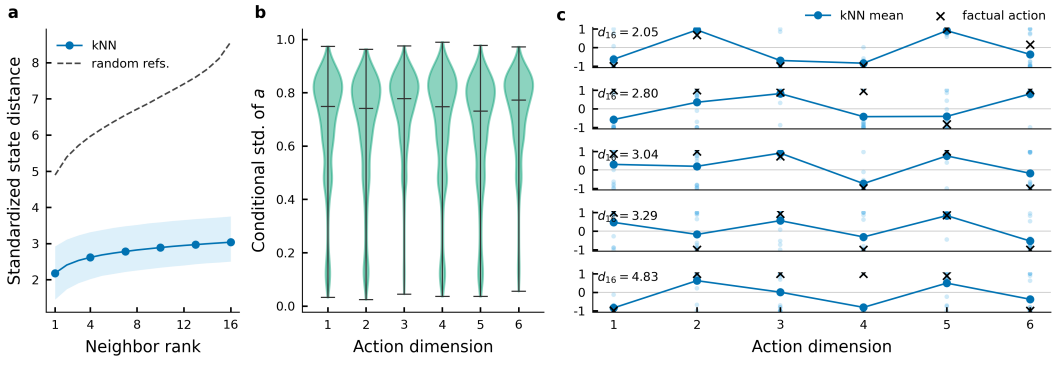


Figure 27: **Reserved diagnostic for empirical action-centering of Walker.** (a) standardized neighbor distance versus neighbor rank; (b) per-action-dimension conditional standard deviation; and (c) kNN action samples, their mean, and the factual action for representative query states.

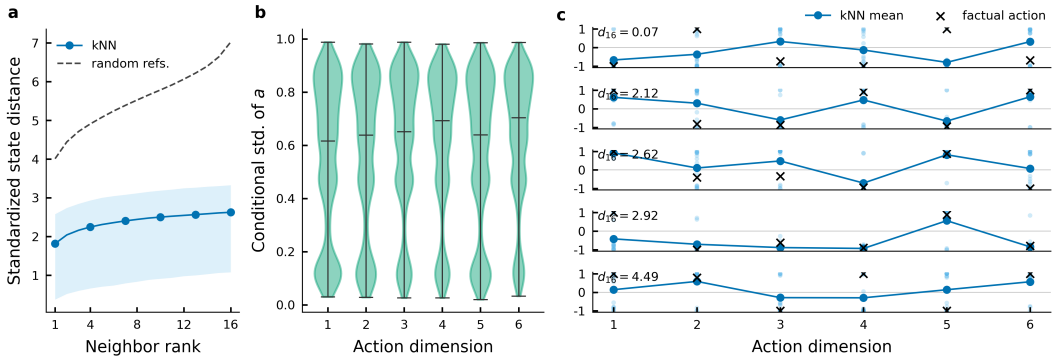


Figure 28: **Reserved diagnostic for empirical action-centering of Cheetah.** (a) standardized neighbor distance versus neighbor rank; (b) per-action-dimension conditional standard deviation; and (c) kNN action samples, their mean, and the factual action for representative query states.

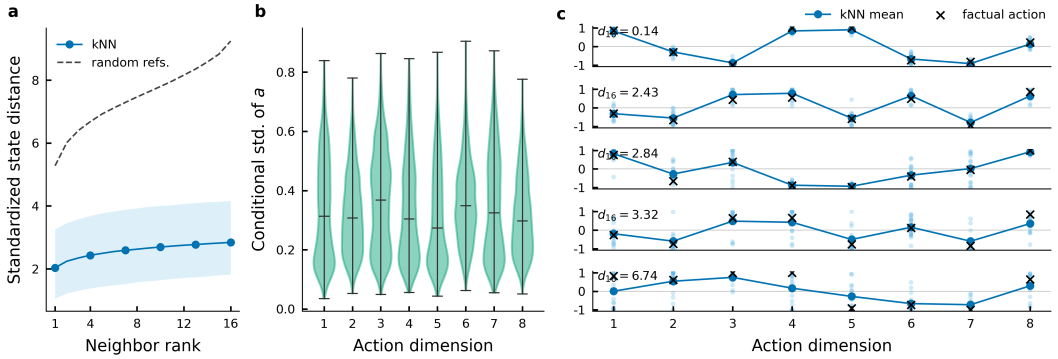


Figure 29: **Reserved diagnostic for empirical action-centering of Antmaze large.** (a) standardized neighbor distance versus neighbor rank; (b) per-action-dimension conditional standard deviation; and (c) kNN action samples, their mean, and the factual action for representative query states.

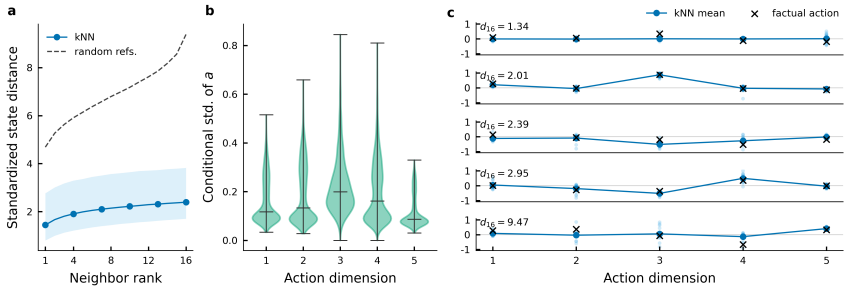


Figure 30: **Reserved diagnostic for empirical action-centering of Cube Single Play.** (a) standardized neighbor distance versus neighbor rank; (b) per-action-dimension conditional standard deviation; and (c) kNN action samples, their mean, and the factual action for representative query states.

G.3.5 Representation learning and calibration

For a positive-lag outcome Y and $X = (S, A)$, the learned delayed component uses the action-centered left feature and marginal-centered right feature. The forward and reverse vector predictors are trained with one-step TD and Polyak targets. After representation learning, an independent

reward-free split is used to estimate feature means and covariance matrices, whiten both sides, and compute a $d \times d$ SVD. The canonical modes are therefore

$$\widehat{f} = f_c T_f, \qquad \widehat{g} = (g_{\text{raw}} - \widehat{\mu}_g) T_g, \qquad \widehat{\Sigma} = \text{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_d). \tag{102}$$

Table 18: **Optimization and representation hyperparameters.** Domain-specific regularization coefficients match the public One-Step FB settings.

Hyperparameter	Walker	Cheetah	AntMaze Large	Cube Single Play
Optimizer	Adam	Adam	Adam	Adam
Adam $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.999, 10^{-8})$	$(0.9, 0.999, 10^{-8})$	$(0.9, 0.999, 10^{-8})$	$(0.9, 0.999, 10^{-8})$
Actor learning rate (One-Step FB aligned)	10^{-4}	10^{-4}	10^{-4}	10^{-4}
Feature learning rate (f, g)	10^{-5}	10^{-5}	10^{-5}	10^{-5}
TD-predictor learning rate (h, v)	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Weight decay / AMSGrad	0/false	0/false	0/false	0/false
Minibatch size	1024	1024	1024	1024
Actor gradient clipping	none	none	none	none
Feature/predictor gradient clipping	global norm 10	global norm 10	global norm 10	global norm 10
Polyak target coefficient τ	0.01	0.01	0.005	0.005
Representation outer iterations	10,000	10,000	10,000	10,000
Forward-TD updates / outer iteration	10	10	10	10
Reverse-TD updates / outer iteration	10	10	10	10
f updates / outer iteration	1	1	1	1
g updates / outer iteration	1	1	1	1
Reverse-TD objective	discounted boundary MSE			
Nested ordering	sequential	sequential	sequential	sequential
Covariance-anchor coefficient	100	100	100	100
Orthonormality/covariance coefficient	0.1	1.0	0	0.3
Forward-head aggregation for actor	mean	mean	min	min
Latent/outcome mixing probability	0.5	0.5	0.5	0.5
Normalize actor score loss by $\mathbb{E}\lvert\widehat{A}\rvert$	yes	yes	yes	yes
BC coefficient λ_{BC}	0	0	0.03	0.3
Reward-weighting temperature τ_{reward}	10	3	10	300
Positive-lag training pairs	10^6	10^6	10^6	10^6
Diagnostic positive-lag pairs	10^5	10^5	10^5	10^5
Calibration-fit snippets	120,000	120,000	120,000	120,000
Calibration-selection snippets	120,000	120,000	120,000	120,000
Calibration horizon	500	500	512	512
Calibration minibatch size	4096	4096	4096	4096
Calibration split	independent	independent	independent	independent
Calibration arithmetic	float64	float64	float64	float64
Ridge multiplier candidates	$\{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$			
Shrinkage multiplier candidates	$\{0, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$			

The covariance ridges for the two sides are selected jointly from the stated grid by the reward-free held-out quadratic loss. The shrinkage multiplier is then selected from its stated grid using the same held-out criterion. Thus a selected value may differ across seeds, but the selection rule and candidate sets are fixed before inspecting downstream task performance. We store the complete sweep and selected values in each calibration artifact.

G.3.6 Reward encoding and actor update

During training, using a set of training rewards $(Y_i, r_i)_{i=1}^N$, reward embeddings follows,

$$\beta_r = \sum_{i=1}^{N} r_i \widehat{g}(Y_i). \tag{103}$$

Unlike the upstream implementation, our primary SUAF variant does not project β_r onto a fixed-radius sphere and does not multiply it by $\widehat{\Sigma}$ before passing it to the actor. Thus both its direction and magnitude are available to the policy. Singular values enter only the SUAF advantage,

$$\widehat{A}_r(s, a) = \widehat{C}_{\pi_b} r(s, a) + \frac{\gamma}{1-\gamma} \widehat{f}(s, a)^\top \widehat{\Sigma} \beta_r. \tag{104}$$

For rewards represented as functions of the stored current state, $\widehat{C}_{\pi_b} r = 0$. For genuinely action-dependent or next-state rewards, the immediate contrast is estimated separately and retained. It is not folded into the positive-lag factorization.

The actor is trained by reparameterized policy-gradient plus behavioral cloning,

$$\mathcal{L}_{\text{actor}} = -\,\text{sg}\left(\frac{1}{\mathbb{E}|\widehat{A}_r|}\right)\mathbb{E}\left[\widehat{A}_r\big(S, \pi_\phi(S, \beta_r)\big)\right] + \lambda_{\text{BC}}\mathbb{E}\left[-\log \pi_\phi(A \mid S, \beta_r)\right]. \tag{105}$$

At evaluation time, each action is produced by one actor forward pass. We do not enumerate action candidates. The upstream fixed-radius latent normalization is retained as a separately labeled ablation; it is not mixed with the raw-β main result.