MIM-DICE: UNSUPERVISED SKILL DISCOVERY FROM ARBITRARY OFF-POLICY EXPERIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised reinforcement learning (URL) aims to acquire reusable behaviors without external supervision for rapid adaptation to downstream tasks. Within URL, skill learning achieves this by discovering skills whose induced state distributions are mutually distinct yet collectively provide broad state coverage. A common way to induce these properties is to maximize the mutual information (MI) between skills and policy-induced states. However, existing methods entail a state distribution mismatch between the off-policy replay data and the current policy. This leads to two critical limitations in practice. First, as we empirically show, their biased estimation of MI impedes effective MI maximization. Second, they strictly rely on replay data collected by skill-conditioned policies, precluding skill learning from arbitrary offline datasets. We introduce Mutual Information Maximization Distribution Correction Estimation (MIM-DICE), a DICE-based method that corrects this distribution mismatch, enabling off-policy unbiased and effective MI maximization. Furthermore, by eliminating the reliance on skill-conditioned data, MIM-DICE enables skill learning directly from arbitrary offline datasets. In online off-policy experiments on tabular environments and Pendulum, MIM-DICE better maximizes MI under the current policy than prior methods. Across six MAZE domains, it learns diverse skills directly from arbitrary offline datasets, achieving higher MI and inducing more clearly separated state-visitation regions. Finally, MIM-DICE outperforms a broad range of state-of-the-art baselines on the URLB benchmark suite.

1 INTRODUCTION

While standard reinforcement learning (RL) typically relies on task-specific rewards, a central goal of the field is to enable agents to adapt efficiently across a wide range of downstream tasks. This gap motivates unsupervised reinforcement learning (URL), which trains agents without external supervision using task-agnostic objectives to facilitate downstream adaptation (Laskin et al., 2021). For example, state-entropy maximization (Liu & Abbeel, 2021b; Seo et al., 2021; Lee et al., 2025) provides robustness to worst-case reward functions (Eysenbach et al., 2022), while novelty-seeking behavior (Pathak et al., 2017; 2019; Burda et al., 2019) may improve adaptation in sparse-reward environments.

We focus on mutual information maximization (MIM) (Eysenbach et al., 2019; Hansen et al., 2020; Sharma et al., 2020; Liu & Abbeel, 2021a; Laskin et al., 2022) between the skill variable and the states induced by the policy. MIM encourages different skills to induce distinct state distributions while collectively covering the state space, yielding a diverse set of reusable behaviors. For this reason, this approach is also known as skill learning.

Although skill learning provides a conceptually appealing objective for URL, existing methods introduce a theoretical mismatch through off-policy training. Specifically, while MIM is defined under the current-policy state occupancies, existing methods use replay-buffer state occupancies. This replacement imposes a practical limitation: estimating replay-buffer state occupancies requires skill-labeled replay data, precluding skill learning from arbitrary unlabeled offline datasets. More fundamentally, even with skill-labeled data, the resulting objective does not generally coincide with the current-policy MI objective. We show in Section 3.1 that this mismatch hinders effective MIM.

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 limitations, we introduce Mutual Information Maximization Distribution Correction Estimation (MIM-DICE), a distribution-correction framework for skill learning from arbitrary offline datasets. Using DICE-based distribution correction, MIM-DICE finds skill-conditioned target occupancies that maximize MI while satisfying the MDP occupancy constraints. This formulation removes the off-policy bias in MI maximization without requiring skill-conditioned data.

We evaluate MIM-DICE in three settings. First, in online off-policy experiments on tabular MDPs and Pendulum, we compare MIM-DICE with existing MIM methods (Eysenbach et al., 2019; Hansen et al., 2020; Liu & Abbeel, 2021a; Laskin et al., 2022; Yang et al., 2023) in terms of policy-induced MI. Compared with prior methods affected by off-policy bias, MIM-DICE achieves the highest policy MI among the evaluated methods. Second, across six maze domains, we evaluate whether MIM-DICE can learn skills from fixed offline datasets with low initial MI. Among the evaluated MIM methods (Liu & Abbeel, 2021a; Laskin et al., 2022) , MIM-DICE is the only method that effectively maximizes policy-induced MI in the fully offline setting. The learned skills exhibit visually distinct state-visitation patterns. Third, we compare MIM-DICE with existing methods, including previous state-of-the-art methods (Laskin et al., 2022; Ying et al., 2026; Bai et al., 2024; Yang et al., 2023), on the URLB (Laskin et al., 2021) downstream tasks. Aggregated across tasks and seeds using RLiable (Agarwal et al., 2021), MIM-DICE achieves the best overall performance. To the best of our knowledge, MIM-DICE is the first method to enable fully offline continuous-skill learning via MIM.

2 PRELIMINARIES

In this section, we introduce our notation and DICE-related background used throughout the paper.

Problem Setup and Notation We consider a discounted Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \rho_0, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s' \mid s, a)$ is the transition kernel, ρ_0 is the initial-state distribution, and $\gamma \in [0, 1)$ is the discount factor. A skill $z \in \mathcal{Z}$ is sampled from a prior $p(z)$, and the agent follows a stationary skill-conditioned policy $\pi(\cdot \mid s, z)$.

For each skill z, we denote by $d_z^\pi(s, a)$ the normalized discounted state-action occupancy induced by π, given by $d_z^\pi(s, a) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_\pi(s_t = s, a_t = a \mid z)$. Its state marginal is $\bar{d}_z^\pi(s) \coloneqq \sum_a d_z^\pi(s, a)$, and the skill-marginalized state occupancy is $\bar{d}^\pi(s) \coloneqq \mathbb{E}_{z\sim p(z)}[\bar{d}_z^\pi(s)]$. A nonnegative skill-conditioned occupancy collection $\{d_z\}$ satisfies the discounted Bellman flow constraints if

$$\sum_a d_z(s', a) = (1-\gamma)\rho_0(s') + \gamma \sum_{s,a} d_z(s, a) P(s' \mid s, a), \qquad \forall s', z. \tag{1}$$

We denote the set of such collections by $\mathcal{D}_{\text{flow}}$.

Additionally, we assume access to an off-policy replay buffer $\mathcal{B}$ of transitions (s, a, s'), optionally accompanied by skill labels z. When no skill labels are available, we refer to $\mathcal{B}$ as an *arbitrary off-policy dataset*. We denote its state-action distribution by $d^{\mathcal{B}}(s, a)$, its state marginal by $\bar{d}^{\mathcal{B}}(s) \coloneqq \sum_a d^{\mathcal{B}}(s, a)$, and, when defined, its skill-conditional state distribution by $\bar{d}_z^{\mathcal{B}}(s)$. We write $(s, a, s') \sim d^{\mathcal{B}}(s, a)P(s' \mid s, a)$ for transitions sampled from the corresponding data distribution. For skill-labeled buffers, we denote the MI between states and skills by $I_{\mathcal{B}}(S; Z)$.

Skill Learning via MIM. MI-based skill learning maximizes $I_\pi(S; Z) = H_\pi(S) - H_\pi(S \mid Z)$. The first term, the state entropy, encourages broad state-space coverage, while the second term, the negative skill-conditioned state entropy, encourages individual skills to induce distinct state distributions. Together, these effects yield mutually distinct skills that collectively cover the state space.

Overview of DICE Consider an optimization problem $\max_{d\in\mathcal{D}_{\text{flow}}} F(d)$. Direct optimization over d requires enforcing the Bellman-flow constraints and thus generally depends on the unknown transition dynamics. DICE-based approaches typically leverage strong duality to eliminate optimization variables and f-divergence regularization to make the remaining terms estimable from off-policy data. As an example, consider a linear occupancy objective $\max_d \langle d, c\rangle$. Adding a concave regularizer gives $\sup_{d\geq 0}\{\langle d, c\rangle - \alpha \mathbb{E}_{d^{\mathcal{B}}}[f(d/d^{\mathcal{B}})]\}$. Letting $w = d/d^{\mathcal{B}}$, Fenchel duality gives

$$\sup_{w\geq 0} \mathbb{E}_{d^{\mathcal{B}}}\left[wc - \alpha f(w)\right] = \alpha \mathbb{E}_{d^{\mathcal{B}}}\left[f^*\left(\frac{c}{\alpha}\right)\right],$$

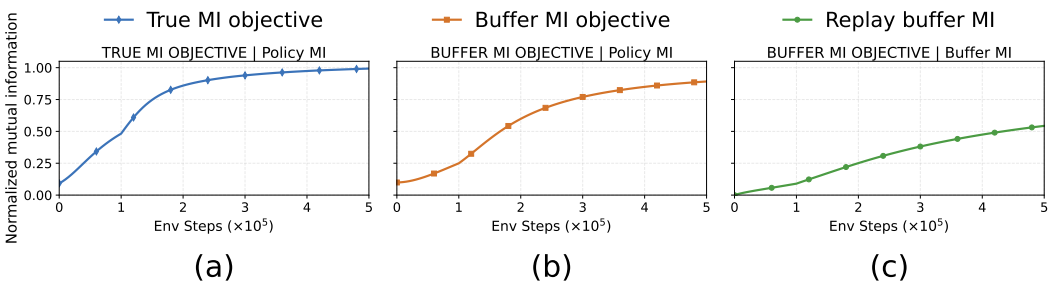


Figure 1: **Replay-buffer MI is not a faithful surrogate for current-policy MI.** In sparse tabular MDPs, we report (a) $I_\pi(S;Z)$ under on-policy MIM, (b) $I_\pi(S;Z)$ under off-policy MIM using the replay-derived objective $J_\mathcal{B}$, and (c) replay-buffer MI $I_\mathcal{B}(S;Z)$ for the same agent as in (b). Off-policy MIM achieves lower policy MI, while $I_\mathcal{B}(S;Z)$ differs substantially from $I_\pi(S;Z)$. Curves show means over 100 random MDP seeds; values are jointly min–max normalized across methods.

where $f^*(y) \coloneqq \sup_{x\geq 0}\{xy - f(x)\}$ is the Fenchel conjugate of f, and $w^* = f^{*\prime}(c/\alpha)$. This makes the target occupancy $\bar{d}$ estimable from off-policy data, motivating concave occupancy objectives in DICE-based optimization. Throughout this paper, we refer to this primal-to-dual transformation as DICE-based dualization, and to any reformulation yielding such an objective as a DICE-based formulation.

3 MIM-DICE

In this section, we derive the MIM-DICE objective and present the resulting method.

3.1 Why Off-Policy MIM Requires Occupancy Correction

We first formalize the replay–policy distribution mismatch discussed in the Introduction. The target MIM objective is

$$I_\pi(S;Z) = H_\pi(S) - H_\pi(S \mid Z) = \mathbb{E}_{z\sim p(z),\, s\sim \bar{d}^\pi_z(\cdot)}\left[-\log \bar{d}^\pi(s) + \log \bar{d}^\pi_z(s)\right].$$

Off-policy MIM methods instead use replay-buffer occupancies $\bar{d}^\mathcal{B}$ and $\bar{d}^\mathcal{B}_z$, corresponding to

$$J_\mathcal{B}(\pi) = \mathbb{E}_{z\sim p(z),\, s\sim \bar{d}^\pi_z}\left[-\log \bar{d}^\mathcal{B}(s) + \log \bar{d}^\mathcal{B}_z(s)\right].$$

The two objectives differ by

$$J_\mathcal{B}(\pi) - I_\pi(S;Z) = D_{\mathrm{KL}}\big(\bar{d}^\pi \,\|\, \bar{d}^\mathcal{B}\big) - \mathbb{E}_{z\sim p(z)}\left[D_{\mathrm{KL}}\big(\bar{d}^\pi_z \,\|\, \bar{d}^\mathcal{B}_z\big)\right],$$

Thus, $J_\mathcal{B}(\pi)$ does not generally coincide with $I_\pi(S;Z)$. We next examine this effect empirically.

We use randomly generated tabular MDPs, where policy- and replay-induced state–skill distributions can be computed exactly. Figure 1 compares (a) policy MI under on-policy MIM, (b) policy MI under off-policy MIM using the replay-derived reward $\bar{r}_\mathcal{B}(s,z) = -\log \bar{d}^\mathcal{B}(s) + \log \bar{d}^\mathcal{B}_z(s)$, and (c) the corresponding replay-buffer MI $I_\mathcal{B}(S;Z)$. Off-policy MIM in (b) consistently attains lower $I_\pi(S;Z)$ than the on-policy counterpart in (a), while $I_\mathcal{B}(S;Z)$ in (c) substantially differs from $I_\pi(S;Z)$. These results show that replay-buffer MI is not a faithful surrogate for current-policy MI, motivating occupancy correction.

3.2 Skill-Conditioned Occupancy Optimization Problem

We now formulate the MI optimization objective to avoid off-policy bias. Instead of optimizing policies using the reward $\bar{r}_\mathcal{B}$ estimated from off-policy data, we directly optimize skill-conditioned occupancy measures.

A natural concern is that optimizing over occupancies may admit distributions that cannot be induced by any policy, in which case it would be impossible to recover a corresponding policy. This concern is resolved by the standard occupancy–policy correspondence. For each skill z, a nonnegative

state–action occupancy d_z is realizable by a stationary Markov policy if and only if it satisfies the discounted Bellman flow constraints Eq. 1. Moreover, for any d_z satisfying Eq. 1, the inducing stationary Markov skill-conditioned policy can be recovered as $\pi_d(a \mid s, z) = d_z(s, a) / \sum_{a'} d_z(s, a')$ whenever $\sum_{a'} d_z(s, a') > 0$, with $\pi_d(\cdot \mid s, z)$ defined arbitrarily otherwise. Thus, any feasible d_z under these constraints can be realized by a stationary Markov policy.

Having established realizability, we return to the MI maximization problem over occupancies. Recall that $\mathcal{D}_{\text{flow}}$ denotes the set of occupancy collections satisfying Eq. 1. For any $\{d_z\} \in \mathcal{D}_{\text{flow}}$, define $\bar{d}_z(s) \coloneqq \sum_a d_z(s, a)$ and $\bar{d}(s) \coloneqq \sum_z p(z) \bar{d}_z(s)$. The corresponding occupancy-induced MI is

$$I_d(S; Z) \coloneqq -\sum_s \bar{d}(s) \log \bar{d}(s) + \sum_z p(z) \sum_s \bar{d}_z(s) \log \bar{d}_z(s). \tag{2}$$

The induced policy π_d satisfies $I_d(S; Z) = I_{\pi_d}(S; Z)$. Since $\mathcal{D}_{\text{flow}}$ is exactly the set of occupancies induced by stationary policies, $\sup_\pi I_\pi(S; Z) = \sup_{\{d_z\} \in \mathcal{D}_{\text{flow}}} I_d(S; Z)$.

Thus, the key requirement is to constrain our optimization variables to lie in $\mathcal{D}_{\text{flow}}$. Given its natural expression in terms of the state marginals $\bar{d}_z(s)$ and $\bar{d}(s)$, the MI optimization can be formulated as:

$$\max_{\{d_z \geq 0,\, \bar{d}_z\}_{z \in \mathcal{Z}},\, d} \quad \underbrace{-\sum_s \bar{d}(s) \log \bar{d}(s)}_{H(S)} + \underbrace{\sum_z p(z) \sum_s \bar{d}_z(s) \log \bar{d}_z(s)}_{-H(S|Z)} - \underbrace{\alpha \sum_z p(z) D_f\big(d_z \,\|\, d^{\mathcal{B}}\big)}_{\text{concave regularizer}} \tag{3}$$

$$\text{s.t.} \quad \sum_a d_z(s', a) = (1 - \gamma) \rho_0(s') + \gamma \sum_{s,a} d_z(s, a) P(s' \mid s, a), \qquad \forall s', z \tag{4}$$

$$\bar{d}_z(s) = \sum_a d_z(s, a), \qquad \forall s, z \tag{5}$$

$$\bar{d}(s) = \sum_z p(z) \sum_a d_z(s, a), \quad \forall s \tag{6}$$

The first two terms in Eq. 3 constitute the MI objective. The remaining f-divergence term acts as a trust-region-like penalty, keeping the target occupancies close to the off-policy data distribution.

Constraint Eq. 4 ensures that each d_z lies in $\mathcal{D}_{\text{flow}}$, while Eq. 5 defines its corresponding state marginal $\bar{d}_z$. Finally, Eq. 6 introduces the skill-marginalized state distribution as an explicit variable and couples the skill-conditioned occupancies through an affine relation, exposing the marginal entropy as a concave function of $\bar{d}$ for the subsequent dual derivation. Note that only the optimization variables d_z are skill-conditioned; $d^{\mathcal{B}}$ may correspond to an *arbitrary off-policy dataset*.

The primal problem seeks valid skill-conditioned occupancies that maximize MI while regularizing distribution shift from the *arbitrary off-policy data*. However, solving it directly requires access to the unknown transition dynamics and explicit target occupancies. In the following subsection, we derive a DICE-based formulation estimable from off-policy transitions, eliminating both requirements.

3.3 Minorization-Maximization for a Difference-of-Convex Objective

$$I_d(S; Z) = \underbrace{-\sum_s \bar{d}(s) \log \bar{d}(s)}_{\text{concave}} + \underbrace{\sum_z p(z) \sum_s \bar{d}_z(s) \log \bar{d}_z(s)}_{\text{convex}} \underbrace{-\alpha \sum_z p(z) D_f(d_z \,\|\, d_D)}_{\text{concave}}.$$

Before deriving the DICE-based formulation, we must first address the nonconcavity of the primal objective. Recall from Preliminaries 2 that DICE-based dualization requires a concave occupancy-measure maximization problem for which standard strong-duality results can be invoked. However, the negative conditional-entropy term renders the objective nonconcave in the occupancy measures, giving the primal a difference-of-convex (DC) structure. Consequently, it does not satisfy the concavity requirement for direct DICE-based dualization. We therefore seek an iterative reformulation whose updates satisfy two properties: (i) each update reduces to a concave occupancy-measure optimization problem amenable to DICE-based dualization, and (ii) the original objective is nondecreasing across iterations.

Minorization–maximization (MM) naturally satisfies these requirements. At iteration k, MM constructs a surrogate $M(d, q)$ that globally lower-bounds original objective $J(d)$ and satisfies the

following properties, with q^k denoting the fixed q at the current iteration and d^{k+1} chosen from $\arg\max_d M(d, q^k)$:

$$q^{k+1} \in \arg\max_q M(d^{k+1}, q), \qquad M(d^{k+1}, q^{k+1}) := \max_q M(d^{k+1}, q) = J(d^{k+1}).$$

These properties imply $J(d^{k+2}) = M(d^{k+2}, q^{k+2}) \geq M(d^{k+2}, q^{k+1}) \geq J(d^{k+1})$ which guarantees requirement (ii). If $M(d, q)$ is additionally concave in d for fixed q, then each update is a concave occupancy-measure optimization problem, satisfying requirement (i).

We now apply this MM construction to the MIM-DICE objective. Specifically, we introduce an auxiliary distribution q to minorize the negative conditional-entropy term. Let $J(\mathbf{d})$ denote the objective in Eq. 3, where $\mathbf{d}$ collects all occupancy variables. Using an auxiliary conditional distribution $q(s \mid z)$, we define the surrogate objective as

$$M(\mathbf{d}, q) = -\sum_s \bar{d}(s) \log \bar{d}(s) + \sum_z p(z) \sum_s \bar{d}_z(s) \log q(s \mid z) - \alpha \sum_z p(z) D_f\big(d_z \,\big\|\, d^{\mathcal{B}}\big). \quad (7)$$

The difference between the original and surrogate objectives is

$$J(\mathbf{d}) - M(\mathbf{d}, q) = \sum_z p(z) D_{\mathrm{KL}}\big(\bar{d}_z \,\big\|\, q(\cdot \mid z)\big) \geq 0. \quad (8)$$

Hence, $M(\mathbf{d}, q)$ globally lower-bounds $J(\mathbf{d})$, with equality when $q(\cdot \mid z) = \bar{d}_z$. Equivalently, $J(\mathbf{d}) = \max_{q:\, q(\cdot|z) \in \Delta(\mathcal{S}),\, \forall z} M(\mathbf{d}, q)$. At iteration k, MM tightens the bound and updates as

$$q^k(\cdot \mid z) = \bar{d}^k_z, \qquad \mathbf{d}^{k+1} \in \arg\max_{\mathbf{d} \in \mathcal{D}_{\mathrm{flow}}} M(\mathbf{d}, q^k). \quad (9)$$

For fixed q^k, the conditional-entropy surrogate is linear in the skill-conditioned occupancies, and $M(\mathbf{d}, q^k)$ is therefore concave in $\mathbf{d}$. The resulting MM updates monotonically improve J, while each occupancy update is amenable to DICE-based dualization.

3.4 DICE-BASED DUALIZATION

Fixing $q^k = \{q^k_z\}_{z \in \mathcal{Z}}$, with $q^k_z \in \Delta(\mathcal{S})$, each MM iteration solves

$$\max_{\{d_z \geq 0\}_{z \in \mathcal{Z}},\, \bar{d}} \quad -\sum_s \bar{d}(s) \log \bar{d}(s) + \sum_z p(z) \sum_{s,a} d_z(s,a) \log q^k_z(s) - \alpha \sum_z p(z) D_f\big(d_z \,\big\|\, d^{\mathcal{B}}\big) \quad (10)$$

$$\text{s.t.} \quad \sum_a d_z(s', a) = (1-\gamma)\rho_0(s') + \gamma \sum_{s,a} d_z(s,a) P(s' \mid s, a), \qquad \forall s', z, \quad (11)$$

$$\bar{d}(s) = \sum_z p(z) \sum_a d_z(s, a), \qquad \forall s. \quad (12)$$

For fixed q^k, the second term in Eq. 10 is linear in the skill-conditioned state marginals. Hence, the objective is concave in the occupancy variables, while the constraints are affine. Nevertheless, solving the problem in primal form would require access to the transition dynamics and explicit optimization over the target occupancies. Our goal is instead to express each MM update using only samples from the off-policy dataset.

We introduce dual variables $\lambda_z(s)$ for the Bellman-flow constraints and $\mu(s)$ for the skill-marginalization constraint. Maximizing the resulting Lagrangian over the primal occupancy variables can be carried out analytically. The occupancy $\bar{d}$ is eliminated using the convex conjugate of negative entropy, while occupancy d_z is eliminated using the convex conjugate f^* of the divergence generator. This yields the following DICE-based dual formulation.

Proposition 1 (DICE dual of the MM subproblem)**.** *For fixed $q^k = \{q^k_z\}_{z \in \mathcal{Z}}$, the MM occupancy problem Eq. 10–Eq. 12 is equivalent to*

$$\min_{\lambda, \mu} \quad \sum_s \exp\big(-1 - \mu(s)\big) + (1-\gamma) \sum_{z \in \mathcal{Z}} \sum_s \lambda_z(s) \rho_0(s)$$
$$+ \alpha \sum_{z \in \mathcal{Z}} p(z) \sum_{s,a} d^{\mathcal{B}}(s,a) f^* \left(\frac{p(z)\big(\log q^k_z(s) + \mu(s)\big) - \lambda_z(s) + \gamma \sum_{s'} P(s' \mid s, a) \lambda_z(s')}{\alpha p(z)} \right). \quad (13)$$

The complete derivation, including the Lagrangian construction and the analytic elimination of the primal occupancies, is provided in Appendix A.

Although Eq. 13 no longer involves the unknown target occupancies, it is not yet directly convenient for sample-based optimization. First, the exponential state sum arising from the marginal-entropy term can be numerically unstable. We therefore replace it with its log-sum-exp counterpart, which preserves the optimal solution as shown in Appendix A.2. Second, the resulting state sum still requires summation over the entire state space. We rewrite it as $\sum_s \exp(-1-\mu(s)) = \mathbb{E}_{s\sim d^{\mathcal{B}}}\left[\frac{\exp(-1-\mu(s))}{\bar{d}^{\mathcal{B}}(s)}\right]$, allowing the term to be estimated from dataset samples. In practice, since $\bar{d}^{\mathcal{B}}(s)$ is unknown, we estimate the dataset state density $\bar{d}^{\mathcal{B}}(s)$ using the density estimator from the URLB implementations of prior state-entropy maximization methods (Liu & Abbeel, 2021b; Lee et al., 2025) that estimate $\bar{d}^{\mathcal{B}}$.

For fixed $\mathbf{q}^k = \{q_z^k\}_{z\in\mathcal{Z}}$, these transformations give

$$\begin{aligned}\min_{\lambda,\mu} \quad & \log \mathbb{E}_{s\sim d^{\mathcal{B}}}\left[\frac{\exp\big(-1-\mu(s)\big)}{\bar{d}^{\mathcal{B}}(s)}\right] + (1-\gamma)\sum_{z\in\mathcal{Z}}\mathbb{E}_{s\sim\rho_0}\left[\lambda_z(s)\right] \\ & + \alpha\sum_{z\in\mathcal{Z}} p(z)\mathbb{E}_{(s,a)\sim d^{\mathcal{B}}}\left[f^*\left(\frac{p(z)\big(\log q_z^k(s)+\mu(s)\big)-\lambda_z(s)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\lambda_z(s')\right]}{\alpha p(z)}\right)\right].\end{aligned} \tag{14}$$

We now return to the auxiliary distributions $\{q_z\}_{z\in\mathcal{Z}}$ used in the MM construction. Let $\mathbf{q} = \{q_z\}_{z\in\mathcal{Z}}$, and let $\mathcal{G}(\lambda,\mu;\mathbf{q})$ denote the objective minimized in Eq. 14, with q_z^k replaced by q_z. As shown in Eq. 8, for any fixed occupancy collection $\mathbf{d}$, the gap between the original objective and its surrogate is $\sum_{z\in\mathcal{Z}} p(z) D_{\mathrm{KL}}\left(\bar{d}_z \,\|\, q_z\right)$. Therefore, maximizing $M(\mathbf{d},\mathbf{q})$ over valid auxiliary distributions $\mathbf{q}$ recovers the original objective, with the optimum attained at $q_z = \bar{d}_z$: $J(\mathbf{d}) = \max_{\{q_z\in\Delta(\mathcal{S})\}_{z\in\mathcal{Z}}} M(\mathbf{d},\mathbf{q})$.

Combining this variational representation with the fixed-$\mathbf{q}$ dual formulation, and enforcing the normalization of each q_z using Lagrange multipliers $\beta(z)$, yields

$$\max_{\{q_z\geq 0\}_{z\in\mathcal{Z}}} \min_{\lambda,\mu,\beta}\left\{\mathcal{G}(\lambda,\mu;\mathbf{q}) - \sum_{z\in\mathcal{Z}}\beta(z)\left(\sum_{s\in\mathcal{S}} q_z(s) - 1\right)\right\}.$$

The remaining normalization constraint on q_z still requires summation over the entire state space. As in the analogous case above, we use the dataset density estimator $\bar{d}^{\mathcal{B}}$ to rewrite the constraint as $\sum_{s\in\mathcal{S}} q_z(s) = \mathbb{E}_{s\sim d^{\mathcal{B}}}\left[\frac{q_z(s)}{\bar{d}^{\mathcal{B}}(s)}\right]$. In practice, q_z is constrained to be strictly positive using a small numerical floor.

Finally, the conditional expectation $\mathbb{E}_{s'\sim P(\cdot|s,a)}[\lambda_z(s')]$ appearing inside f^* is replaced by the single-sample estimate $\lambda_z(s')$, where $s' \sim P(\cdot \mid s,a)$. This yields the following sample-based MIM-DICE objective:

$$\begin{aligned}\sup_{\{q_z>0\}_{z\in\mathcal{Z}}} \min_{\lambda,\mu,\beta} \quad & \log \mathbb{E}_{s\sim d^{\mathcal{B}}}\left[\frac{\exp\big(-1-\mu(s)\big)}{\bar{d}^{\mathcal{B}}(s)}\right] + (1-\gamma)\sum_{z\in\mathcal{Z}}\mathbb{E}_{s\sim\rho_0}\left[\lambda_z(s)\right] \\ & + \alpha\sum_{z\in\mathcal{Z}} p(z)\mathbb{E}_{(s,a,s')\sim d^{\mathcal{B}}}\left[f^*\left(\frac{p(z)\big(\log q_z(s)+\mu(s)\big)-\lambda_z(s)+\gamma\lambda_z(s')}{\alpha p(z)}\right)\right] \\ & - \sum_{z\in\mathcal{Z}}\beta(z)\left[\mathbb{E}_{s\sim d^{\mathcal{B}}}\left[\frac{q_z(s)}{\bar{d}^{\mathcal{B}}(s)}\right] - 1\right].\end{aligned} \tag{15}$$

Under stochastic dynamics, replacing the conditional expectation inside f^* by a single transition sample is generally biased because f^* is nonlinear. In such cases, the method described in the Appendix E can be used. Under deterministic dynamics, this replacement introduces no bias.

Recall that $d^{\mathcal{B}}$ need not be skill-conditioned. MIM-DICE only requires density ratios between the skill-conditioned occupancies d_z and $d^{\mathcal{B}}$, which allows skill learning from *arbitrary off-policy data*.

Policy Extraction. After training, MIM-DICE extracts an policy using the I-projection (Lee et al., 2021):

$$\min_{\pi} D_{\mathrm{KL}}\big(\bar{d}^{\mathcal{B}}(s)p(z)\pi(a \mid s,z) \,\big\|\, \bar{d}^{\mathcal{B}}(s)p(z)\pi^*(a \mid s,z)\big) = \mathbb{E}_{\substack{s\sim\bar{d}^{\mathcal{B}},\, z\sim p(z) \\ a\sim\pi}}\left[-\log w(s,a,z) + D_{\mathrm{KL}}(\pi(\cdot \mid s,z) \,\|\, \pi_D(\cdot \mid s))\right] + C. \tag{16}$$

Here, $w(s,a,z) = d_z(s,a)/d^{\mathcal{B}}(s,a)$ is the target-to-buffer occupancy ratio, recoverable from the networks learned by MIM-DICE. For each $s \sim \bar{d}^{\mathcal{B}}$ and randomly sampled $z \sim p(z)$, the objective trains the current policy to mimic π^*, which induces d_z. Implementation details and pseudocode are provided in Appendix C.

4 Related Work

Since MI admits two equivalent decompositions, $H(S) - H(S \mid Z)$ and $H(Z) - H(Z \mid S)$, MIM-based URL methods can be broadly categorized into the former, exemplified by APS (Liu & Abbeel, 2021a), and the latter, exemplified by VISR (Hansen et al., 2020). Empirically, the former tends to yield better downstream performance by explicitly rewarding state coverage (Liu & Abbeel, 2021a; Laskin et al., 2022). Variants of these objectives include CIC (Laskin et al., 2022), which maximizes a lower bound on $H(S) - H(S \mid Z)$ using a contrastive discriminator $q(s \mid z)$, and DIAYN (Eysenbach et al., 2019), which applies SAC to maximize $H(Z) - H(Z \mid S) + H(A \mid S, Z)$. However, these methods incur bias when off-policy transitions are used to construct intrinsic rewards.

DOI (Vlastelica et al., 2024) avoids this bias by maximizing $H(Z) - H(Z \mid S)$ with DICE, but targets offline skill learning from expert demonstrations rather than URL. In our experiments, its online variant frequently collapsed to the global minimum of MI, a failure mode also reported in the original paper for the offline setting. We therefore exclude DOI from direct comparison in the main text and report the online results in the Appendix D.2.

5 Experiments

We evaluate MIM-DICE along three dimensions:

Q1. Occupancy Correction: Does correction improve optimization of $I_\pi(S;Z)$?

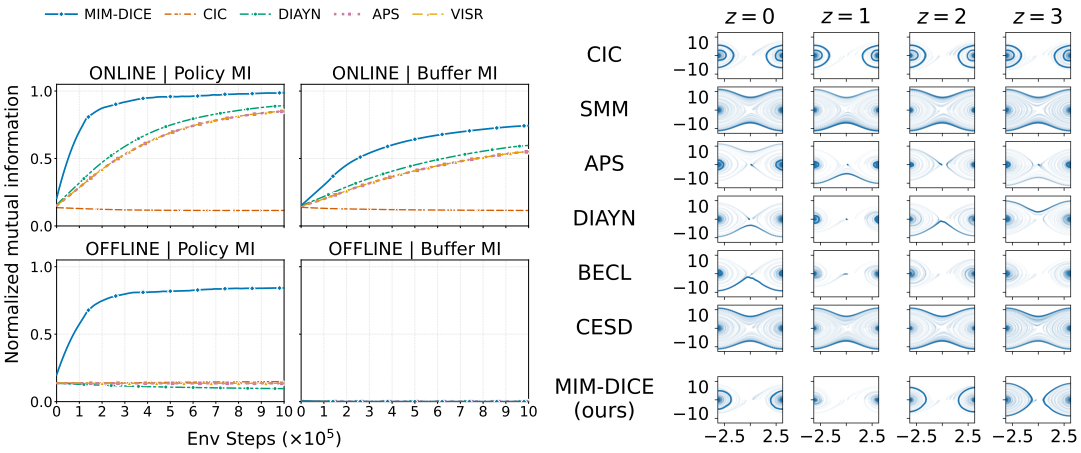
Q2. Offline Skill Discovery: Can MIM-DICE learn skills from arbitrary offline data?

Q3. Downstream Adaptation: Do these advantages translate to downstream performance?

Online Off-Policy Evaluation To address Q1, we evaluate whether occupancy correction improves policy-induced MI maximization in online off-policy skill learning. We consider tabular MDPs, where policy occupancies and MI can be computed exactly, and Pendulum, to evaluate whether the benefit extends to continuous control.

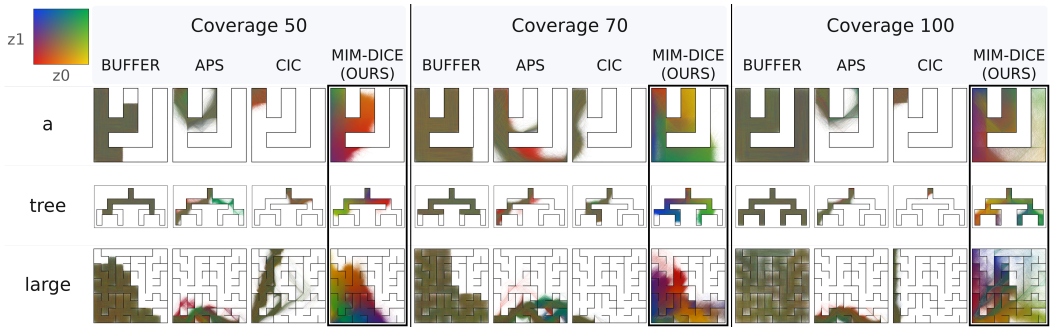
For the tabular experiments, we use randomly generated MDPs with 20 states, 5 actions, and 8 skills, and train each method for 1M environment steps over 100 random seeds. We compare MIM-DICE against DIAYN(Eysenbach et al., 2019), VISR(Hansen et al., 2020), APS(Liu & Abbeel, 2021a), and CIC(Laskin et al., 2022), representative off-policy MIM methods discussed in Section 4. As our primary metric, we exactly compute the policy-induced MI $I_\pi(S;Z)$ throughout training; we additionally report the replay-buffer MI $I_{\mathcal{B}}(S;Z)$ to examine whether the replay distribution faithfully tracks the current-policy objective.

For Pendulum, we compare MIM-DICE with CIC, SMM, APS, DIAYN, BECL, and CeSD (Laskin et al., 2022; Lee et al., 2019; Liu & Abbeel, 2021a; Eysenbach et al., 2019; Yang et al., 2023; Bai et al., 2024) using their official URLB or author-released implementations. We train each method online for 200K environment steps with four one-hot skills. The two-dimensional state space allows us to directly visualize skill-conditioned state visitation and, after discretization, compute the approximate policy-induced mutual information $\hat{I}_\pi(S;Z)$.



(a) Tabular

(b) Pendulum



(c) Maze

Table 1: Results on the Pendulum and MAZE experiments. Statistically best-performing results are bolded.

Method	$\widehat{I}(S;Z)$ (↑)
CIC	0.13 ± 0.05
SMM	0.17 ± 0.06
APS	0.60 ± 0.12
DIAYN	0.66 ± 0.08
BECL	0.37 ± 0.09
CESD	0.02 ± 0.01
MIM-DICE (Ours)	$\mathbf{0.94 \pm 0.05}$

(a) Pendulum

Coverage (%)	**Metric**	**MIM-DICE (Ours)**	**APS**	**CIC**
50	Out-Buffer Rate (%, ↓)	$\mathbf{10.24 \pm 1.71}$	43.52 ± 2.65	23.43 ± 3.38
	In-Buffer Coverage (%, ↑)	$\mathbf{84.2 \pm 8.2}$	64.4 ± 1.3	55.6 ± 2.2
	$\widehat{I}(S;Z)$ (↑)	$\mathbf{2.143 \pm 0.285}$	0.428 ± 0.072	0.096 ± 0.019
70	Out-Buffer Rate (%, ↓)	$\mathbf{10.74 \pm 3.02}$	40.30 ± 1.33	$\mathbf{10.14 \pm 4.29}$
	In-Buffer Coverage (%, ↑)	$\mathbf{80.2 \pm 4.1}$	56.5 ± 1.8	43.6 ± 2.4
	$\widehat{I}(S;Z)$ (↑)	$\mathbf{2.091 \pm 0.163}$	0.302 ± 0.075	0.068 ± 0.006
100	Out-Buffer Rate (%, ↓)	**0.0**	**0.0**	**0.0**
	In-Buffer Coverage (%, ↑)	$\mathbf{80.4 \pm 4.8}$	41.2 ± 2.7	23.0 ± 2.9
	$\widehat{I}(S;Z)$ (↑)	$\mathbf{2.101 \pm 0.188}$	0.393 ± 0.062	0.056 ± 0.012

(b) MAZE

Across both tabular MDPs and Pendulum, MIM-DICE consistently achieves the highest policy-induced MI, supporting the effectiveness of occupancy correction under off-policy training (Figure 2a and Table 1a). In the tabular experiments, the replay-buffer MI of prior methods follows dynamics different from their policy-induced MI, consistent with the replay–policy mismatch identified earlier. On Pendulum, the higher MI is further reflected in clearly separated skill-conditioned state-visitation regions (Figure 2b).

Offline Skill Discovery. To address Q2, we test whether MIM-DICE can create skill structure using only fixed data collected independently of skills. We first consider the same tabular MDPs with a dataset collected by a uniformly random policy, providing a controlled test of fully offline MI maximization against the same baselines as in the online setting.

We then evaluate continuous-skill discovery across six MAZE domains with varying dataset coverage. Each dataset is collected independently of skills, after which transitions are randomly assigned two-dimensional continuous skills z, making S and Z independent. We train for 500K steps using

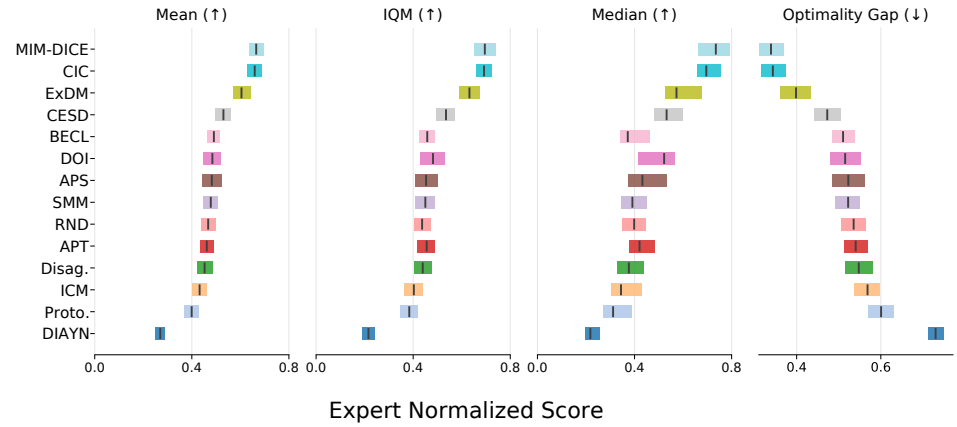


Figure 3: URLB downstream adaptation performance over 10 seeds using RLiable.

only the fixed dataset and compare against APS and CIC, which support continuous skills and optimize $H(S) - H(S \mid Z)$. For visualization, each z is mapped to a fixed color, such that successful skill discovery produces spatially organized skill-conditioned visitation regions.

We discretize the state space to estimate the policy-induced $\hat{I}(S; Z)$. To distinguish meaningful skill discovery from narrow or out-of-distribution solutions, we additionally measure in-buffer coverage, the fraction of buffer-supported states visited by the learned policies, and out-buffer visitation, the fraction of policy visits to states outside the dataset support.

Across both tabular MDPs and MAZE, MIM-DICE is the only evaluated method that effectively increases policy-induced MI under fully offline training. In MAZE, this increase in MI is visually reflected by the emergence of spatially organized skill-conditioned regions from the randomly mixed skill assignments in the offline datasets (Figure 2c). MIM-DICE further covers roughly 80% of the buffer-supported state space while maintaining low out-buffer visitation (Table 1b), indicating that it discovers distinct skills across a broad portion of the available data support rather than through narrow or out-of-distribution behaviors.

Downstream Adaptation. To address Q3, we evaluate MIM-DICE on URLB (Laskin et al., 2021), which measures downstream adaptation after unsupervised pretraining, following the standard benchmark protocol. We compare against prior and recent state-of-the-art URL methods, including CIC, ExDM (Ying et al., 2026),and CeSD under the same evaluation setting. Following the recommended evaluation protocol, we aggregate performance across downstream tasks and random seeds using RLiable (Agarwal et al., 2021). As shown in Figure 3, MIM-DICE achieves the strongest overall downstream performance, indicating that its improvements in skill discovery translate into more effective downstream adaptation rather than merely higher MI.

6 Conclusion

We introduced MIM-DICE, a distribution-correction framework for mutual-information-based skill learning. We showed that the distribution mismatch induced by off-policy training can hinder the maximization of policy-induced MI, and that correcting this mismatch leads to more effective MI maximization. By removing the reliance on skill-conditioned replay data, MIM-DICE further enables fully offline continuous-skill learning directly from arbitrary offline datasets. Across online off-policy, fully offline, and downstream evaluations, MIM-DICE consistently achieved higher policy-induced MI, learned more clearly separated skills, and translated improved skill learning into stronger downstream performance.

Limitation A limitation of MIM-DICE is that the underlying MI maximization problem is nonconcave and therefore does not, in general, admit guarantees of global optimality. The resulting max-min formulation may also pose optimization challenges, although we observed stable behavior in practice. Still, our results highlight the importance of accounting for off-policy distribution mismatch for faithful MI maximization and skill learning from both online replay and offline datasets.

AI USE STATEMENT

We used generative AI tools for manuscript editing, code development, and occasional checks of mathematical derivations written by the authors. The theoretical formulation and interpretation of experimental results were developed by the authors. All AI-assisted work was reviewed by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies unsupervised reinforcement learning in simulated environments. Our experiments use synthetic or benchmark transition data and do not involve human subjects or personally identifiable information. We do not identify specific ethical concerns arising from the experiments conducted in this work. However, applying reinforcement learning methods to real-world systems may introduce safety considerations that are not captured by the benchmarks studied here.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce both the theoretical and empirical results of this work. Section 3 presents the MIM-DICE formulation, and Appendix A provides the detailed derivation of the sample-based objective. Appendix B.1 describes the experimental settings, implementation details, datasets, baselines, and hyperparameters. Appendix C.1 further describes the practical optimization and policy-extraction procedures. Code for reproducing the experiments is available at `https://anonymous.4open.science/r/mimdice-anonymous-4ED5/README.md`.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html`.

Chenjia Bai, Rushuai Yang, Qiaosheng Zhang, Kang Xu, Yi Chen, Ting Xiao, and Xuelong Li. Constrained ensemble exploration for unsupervised skill discovery. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL `https://proceedings.mlr.press/v235/bai24d.html`.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1lJJnR5Ym`.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=SJx63jRqFm`.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. The information geometry of unsupervised reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Steven Hansen, Will Dabney, Andre Barreto, Tom Van de Wiele, David Warde-Farley, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations (ICLR)*, 2020. URL `https://openreview.net/forum?id=BJeAHkrYDS`.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/091d584fced301b442654dd8c23b3fc9-Abstract-round2.html`.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Contrastive intrinsic control for unsupervised reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/debf482a7dbdc401f9052dbe15702837-Abstract-Conference.html.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/lee21f.html.

Jongmin Lee, Meiqi Sun, and Pieter Abbeel. Semdice: Off-policy state entropy maximization via stationary distribution correction estimation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=rJ5g8ueQaI.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric P. Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *CoRR, abs/1906.05274*, 2019.

Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021a. URL https://proceedings.mlr.press/v139/liu21b.html.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/99bf3d153d4bf67d640051a1af322505-Abstract.html.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *ICML*, 2019.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=HJgLZR4KvH.

Marin Vlastelica, Jin Cheng, Georg Martius, and Pavel Kolev. Offline diversity maximization under imitation constraints. In *Reinforcement Learning Conference (RLC) 2024*, 2024. URL https://openreview.net/forum?id=MwpwVMVhH4.

Rushuai Yang, Chenjia Bai, Hongyi Guo, Siyuan Li, Bin Zhao, Zhen Wang, Peng Liu, and Xuelong Li. Behavior contrastive learning for unsupervised skill discovery. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/yang23a.html.

Chengyang Ying, Huayu Chen, Xinning Zhou, Zhongkai Hao, Hang Su, and Jun Zhu. Exploratory diffusion model for unsupervised reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=k0Kb1ynFbt.

A Detailed Derivation of MIM-DICE

In this section, we provide the detailed derivation of the sample-based MIM-DICE objective in Eq. equation 15. The derivation proceeds in three steps. First, for a fixed auxiliary distribution q^k, we dualize the concave occupancy optimization problem in Eq. equation 10–equation 12 so that

the unknown target occupancies and transition dynamics do not need to be optimized explicitly. Second, we rewrite the remaining state-space summations as expectations over the off-policy data distribution. Finally, we optimize the auxiliary distributions $\{q_z\}$ together with the dual variables, which yields the sample-based saddle-point objective used by MIM-DICE.

A.1 DICE Dualization of the MM Subproblem

We first consider one MM iteration with the auxiliary distributions $q^k = \{q_z^k\}_{z\in\mathcal{Z}}$ fixed. The purpose of this step is to eliminate the target occupancies $\{d_z\}$ from the optimization. Although the primal problem has a clear interpretation in terms of skill-conditioned occupancies, solving it directly would require explicit access to the transition dynamics P. We therefore impose the Bellman-flow constraints through dual variables and analytically maximize out the primal occupancy variables.

For fixed q^k, Eq. 10– 12 is

$$\max_{\{d_z\geq 0\}_{z\in\mathcal{Z}},\,\bar{d}} \quad -\sum_s \bar{d}(s)\log\bar{d}(s) + \sum_z p(z)\sum_{s,a} d_z(s,a)\log q_z^k(s) - \alpha\sum_z p(z) D_f(d_z\|d^B)$$

s.t.

$$\sum_a d_z(s',a) = (1-\gamma)\rho_0(s') + \gamma\sum_{s,a} d_z(s,a)P(s'\mid s,a), \qquad \forall s', z, \tag{17}$$

$$\bar{d}(s) = \sum_z p(z)\sum_a d_z(s,a), \qquad \forall s. \tag{18}$$

Lagrangian formulation. The primal problem above is a constrained maximization over the skill-conditioned occupancies. Our goal is to remove these explicit occupancy constraints and obtain a formulation that can eventually be optimized using only off-policy samples.

To this end, we introduce dual variables $\lambda_z(s)$ for the Bellman-flow constraints in Eq. equation 17 and $\mu(s)$ for the skill-marginalization constraint in Eq. equation 18. Under the standard strong-duality conditions for the fixed-q^k concave subproblem, the constrained primal can be equivalently written as the max–min saddle-point problem

$$\max_{\{d_z\geq 0\}_{z\in\mathcal{Z}},\,\bar{d}}\ \min_{\lambda,\mu}\ \mathcal{L}(d,\bar{d},\lambda,\mu;q^k). \tag{19}$$

The corresponding Lagrangian is

$$\begin{aligned}
\mathcal{L}(d,\bar{d},\lambda,\mu;q^k) = &-\sum_s \bar{d}(s)\log\bar{d}(s) + \sum_z p(z)\sum_{s,a} d_z(s,a)\log q_z^k(s) \\
&- \alpha\sum_z p(z)\sum_{s,a} d^B(s,a) f\left(\frac{d_z(s,a)}{d^B(s,a)}\right) \\
&+ \sum_z\sum_{s'}\lambda_z(s')\left[(1-\gamma)\rho_0(s') + \gamma\sum_{s,a} d_z(s,a)P(s'\mid s,a) - \sum_a d_z(s',a)\right] \\
&+ \sum_s \mu(s)\left[\sum_z p(z)\sum_a d_z(s,a) - \bar{d}(s)\right].
\end{aligned} \tag{20}$$

The primal variables now appear only through $\bar{d}(s)$ and $d_z(s,a)$, and both can be maximized out analytically. We consider these two maximizations separately.

For each state s, the terms involving $\bar{d}(s)$ are

$$-\bar{d}(s)\log\bar{d}(s) - \mu(s)\bar{d}(s).$$

Therefore,

$$\sup_{\bar{d}(s)\geq 0}\left\{-\bar{d}(s)\log\bar{d}(s) - \mu(s)\bar{d}(s)\right\} = \exp\{-1-\mu(s)\}, \tag{21}$$

with

$$\bar{d}^\star(s) = \exp\{-1 - \mu(s)\}.$$

Thus, the marginal-entropy term can be represented entirely through the dual function μ.

We next eliminate the skill-conditioned occupancies $\{d_z\}$. To this end, we express each target occupancy relative to the off-policy data distribution through the density ratio

$$w_z(s,a) := \frac{d_z(s,a)}{d^B(s,a)}. \tag{22}$$

To simplify the terms associated with $d_z(s,a)$, we define

$$e_{\lambda,\mu,q^k}(s,a,z) := p(z)\big(\log q_z^k(s) + \mu(s)\big) - \lambda_z(s) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\lambda_z(s')\right]. \tag{23}$$

This quantity collects the coefficient of the skill-conditioned occupancy $d_z(s,a)$ in the Lagrangian, and can be viewed as a Bellman-like residual determined by the dual variables (λ, μ) and the fixed auxiliary distribution q^k.

Using $d_z(s,a) = d^B(s,a) w_z(s,a)$, the terms involving the skill-conditioned occupancies can be written as

$$\sum_z \sum_{s,a} d^B(s,a) \left[w_z(s,a) e_{\lambda,\mu,q^k}(s,a,z) - \alpha p(z) f\big(w_z(s,a)\big) \right]. \tag{24}$$

For each (s,a,z),

$$\sup_{w_z(s,a) \geq 0} \left\{ w_z(s,a) e_{\lambda,\mu,q^k}(s,a,z) - \alpha p(z) f\big(w_z(s,a)\big) \right\} = \alpha p(z) f^* \left(\frac{e_{\lambda,\mu,q^k}(s,a,z)}{\alpha p(z)} \right).$$

Therefore, the maximization over $\{d_z\}$ is reduced to a function of (λ, μ, q^k) evaluated under the off-policy distribution d^B, with no explicit target occupancy remaining.

Combining the maximization over $\bar{d}$ and $\{d_z\}$, for fixed $q^k = \{q_z^k\}_{z \in \mathcal{Z}}$, the resulting dual problem is

$$\min_{\lambda,\mu} \quad \sum_s \exp\{-1-\mu(s)\} + (1-\gamma) \sum_{s,z} \lambda_z(s) \rho_0(s) + \alpha \sum_{s,a,z} p(z) d^B(s,a) f^* \left(\frac{e_{\lambda,\mu,q^k}(s,a,z)}{\alpha p(z)} \right), \tag{25}$$

where

$$e_{\lambda,\mu,q^k}(s,a,z) := p(z)\big(\log q_z^k(s) + \mu(s)\big) - \lambda_z(s) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\lambda_z(s')\right]. \tag{26}$$

This recovers the fixed-q^k DICE dual formulation in Eq. equation 13. The target occupancies no longer appear explicitly, but the first term still requires a summation over the entire state space and involves an exponential of the dual function μ. We next introduce a LogSumExp stabilization for this term before rewriting the objective in a sample-based form.

A.2 LogSumExp Stabilization

The first term in Eq. equation 25,

$$\sum_s \exp\{-1 - \mu(s)\}, \tag{27}$$

can be numerically unstable when μ is parameterized by a neural network. We therefore replace this exponential state sum with its LogSumExp counterpart and show below that the resulting formulation preserves an optimizer of the original objective up to a constant shift of the dual variables.

Define

$$Z(\mu) := \sum_s \exp\{-\mu(s)\}. \tag{28}$$

Then the original exponential term can be written as

$$\sum_s \exp\{-1-\mu(s)\} = \frac{Z(\mu)}{e}. \tag{29}$$

For fixed q, let $\mathcal{G}_{\exp}(\lambda, \mu; q)$ denote the dual objective containing the original term $Z(\mu)/e$, and let $\mathcal{G}_{\text{lse}}(\lambda, \mu; q)$ denote the corresponding objective in which this term is replaced by $\log Z(\mu)$. The LogSumExp term used in Eq. equation 14 is

$$\log \sum_s \exp\{-1-\mu(s)\} = \log Z(\mu) - 1, \tag{30}$$

which differs from $\log Z(\mu)$ only by an additive constant and therefore has the same minimizers.

For any $X > 0$,

$$\log X \le \frac{X}{e}, \tag{31}$$

with equality if and only if $X = e$. Applying this inequality to $Z(\mu)$ gives

$$\log Z(\mu) \le \frac{Z(\mu)}{e} = \sum_s \exp\{-1-\mu(s)\}. \tag{32}$$

Therefore,

$$\mathcal{G}_{\text{lse}}(\lambda, \mu; q) \le \mathcal{G}_{\exp}(\lambda, \mu; q) \tag{33}$$

for every (λ, μ), with equality whenever $Z(\mu) = e$.

We next show that this equality condition can always be satisfied without changing the value of the stabilized objective. For any constant $c \in \mathbb{R}$, define the shifted dual variables

$$\mu_c(s) := \mu(s) + c, \tag{34}$$

$$\lambda_{z,c}(s) := \lambda_z(s) + \frac{p(z)c}{1-\gamma}. \tag{35}$$

Recall the residual

$$\begin{aligned} e_{\lambda,\mu,q}(s,a,z) :=& p(z)\big(\log q_z(s) + \mu(s)\big) - \lambda_z(s) \\ &+ \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\lambda_z(s')\right]. \end{aligned} \tag{36}$$

Under the shift in Eq. equation 34–equation 35, we have

$$\begin{aligned} & e_{\lambda_c,\mu_c,q}(s,a,z) \\ &= p(z)\big(\log q_z(s) + \mu(s) + c\big) - \left(\lambda_z(s) + \frac{p(z)c}{1-\gamma}\right) \\ &\quad + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\lambda_z(s') + \frac{p(z)c}{1-\gamma}\right] \\ &= e_{\lambda,\mu,q}(s,a,z) + p(z)c - \frac{p(z)c}{1-\gamma} + \frac{\gamma p(z)c}{1-\gamma} \\ &= e_{\lambda,\mu,q}(s,a,z). \end{aligned} \tag{37}$$

Hence, the Fenchel-conjugate term in the dual objective is invariant under this shift.

The LogSumExp term changes as

$$\begin{aligned} \log Z(\mu_c) &= \log \sum_s \exp\{-(\mu(s) + c)\} \\ &= \log Z(\mu) - c. \end{aligned} \tag{38}$$

On the other hand, the initial-state term becomes

$$\begin{aligned}
&(1-\gamma)\sum_z\sum_s \rho_0(s)\lambda_{z,c}(s) \\
&= (1-\gamma)\sum_z\sum_s \rho_0(s)\lambda_z(s) + (1-\gamma)\sum_z\sum_s \rho_0(s)\frac{p(z)c}{1-\gamma} \\
&= (1-\gamma)\sum_z\sum_s \rho_0(s)\lambda_z(s) + c,
\end{aligned} \tag{39}$$

where we used $\sum_s \rho_0(s) = 1$ and $\sum_z p(z) = 1$.

Therefore, the $-c$ change in the LogSumExp term is exactly canceled by the $+c$ change in the initial-state term, while the Fenchel-conjugate term remains unchanged. Consequently,

$$\mathcal{G}_{\text{lse}}(\lambda_c, \mu_c; q) = \mathcal{G}_{\text{lse}}(\lambda, \mu; q). \tag{40}$$

Now let $(\lambda^\star, \mu^\star)$ be an optimizer of the stabilized objective for fixed q, and choose

$$c^\star := \log Z(\mu^\star) - 1. \tag{41}$$

Define

$$\widetilde{\mu}(s) := \mu^\star(s) + c^\star, \quad \widetilde{\lambda}_z(s) := \lambda_z^\star(s) + \frac{p(z)c^\star}{1-\gamma}. \tag{42}$$

Then

$$Z(\widetilde{\mu}) = \sum_s \exp\{-(\mu^\star(s) + c^\star)\} = e^{-c^\star} Z(\mu^\star) = e. \tag{43}$$

Thus, the inequality in Eq. equation 32 is tight at $(\widetilde{\lambda}, \widetilde{\mu})$, and therefore

$$\mathcal{G}_{\exp}(\widetilde{\lambda}, \widetilde{\mu}; q) = \mathcal{G}_{\text{lse}}(\widetilde{\lambda}, \widetilde{\mu}; q). \tag{44}$$

Moreover, by the shift invariance in Eq. equation 40,

$$\mathcal{G}_{\text{lse}}(\widetilde{\lambda}, \widetilde{\mu}; q) = \mathcal{G}_{\text{lse}}(\lambda^\star, \mu^\star; q). \tag{45}$$

Since $(\lambda^\star, \mu^\star)$ minimizes the stabilized objective and $\mathcal{G}_{\text{lse}} \le \mathcal{G}_{\exp}$ pointwise, for any (λ, μ) we have

$$\begin{aligned}
\mathcal{G}_{\exp}(\widetilde{\lambda}, \widetilde{\mu}; q) &= \mathcal{G}_{\text{lse}}(\widetilde{\lambda}, \widetilde{\mu}; q) \\
&= \mathcal{G}_{\text{lse}}(\lambda^\star, \mu^\star; q) \\
&\le \mathcal{G}_{\text{lse}}(\lambda, \mu; q) \\
&\le \mathcal{G}_{\exp}(\lambda, \mu; q).
\end{aligned} \tag{46}$$

Hence, $(\widetilde{\lambda}, \widetilde{\mu})$ is also an optimizer of the original exponential-sum objective.

Therefore, minimizing the LogSumExp-stabilized objective and applying the constant shift above recovers an optimizer of the original exponential-sum formulation. Since the LogSumExp term appearing in Eq. equation 14 differs from $\log Z(\mu)$ only by an additive constant, the same conclusion holds for the objective used in MIM-DICE. The stabilized dual objective still contains explicit summations over the state space. Our next goal is to express these terms as expectations under the dataset distribution, so that the objective can be optimized without enumerating states.

Additionally, assuming $\bar{d}^B(s) > 0$ on the relevant state support,

$$\begin{aligned}
\log \sum_s \exp\{-1 - \mu(s)\} &= \log \sum_s \bar{d}^B(s) \frac{\exp\{-1-\mu(s)\}}{\bar{d}^B(s)} \\
&= \log \mathbb{E}_{s\sim\bar{d}^B}\left[\frac{\exp\{-1-\mu(s)\}}{\bar{d}^B(s)}\right].
\end{aligned} \tag{47}$$

A.3 Optimizing the Auxiliary Distributions

So far, q^k has been treated as fixed, corresponding to one MM occupancy update. We now return to the variational representation introduced in Section 3.3. Recall that

$$J(d) = \max_{\{q_z \in \Delta(\mathcal{S})\}_{z \in \mathcal{Z}}} M(d, q),$$

and the bound is tight when $q_z = \bar{d}_z$. Therefore, rather than explicitly alternating between exact MM updates, we optimize q jointly with the dual variables.

To enforce

$$\sum_s q_z(s) = 1$$

for every skill, we introduce Lagrange multipliers $\beta(z)$. This gives the saddle-point formulation

$$\sup_{\{q_z > 0\}_{z \in \mathcal{Z}}} \min_{\lambda, \mu, \beta} \quad G(\lambda, \mu; q) - \sum_z \beta(z) \left(\sum_s q_z(s) - 1 \right). \tag{48}$$

The normalization term can again be expressed using the dataset state marginal:

$$\sum_s q_z(s) = \mathbb{E}_{s \sim \bar{d}^B} \left[\frac{q_z(s)}{\bar{d}^B(s)} \right]. \tag{49}$$

The only remaining quantity that cannot be computed directly from individual dataset transitions is

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}[\lambda_z(s')]$$

inside f^*. Given a sampled transition (s, a, s'), we replace this conditional expectation with the single-sample estimate $\lambda_z(s')$. Together with the importance-sampling identities above, this produces the objective that is optimized in practice:

$$\begin{aligned} \sup_{\{q_z > 0\}_{z \in \mathcal{Z}}} \min_{\lambda, \mu, \beta} \quad & \log \mathbb{E}_{s \sim \bar{d}^B} \left[\frac{\exp\{-1 - \mu(s)\}}{\bar{d}^B(s)} \right] \\ & + (1 - \gamma) \sum_z \mathbb{E}_{s \sim \rho_0}[\lambda_z(s)] \\ & + \alpha \sum_z p(z) \mathbb{E}_{(s,a,s') \sim d^B} \left[f^* \left(\frac{p(z)\big(\log q_z(s) + \mu(s)\big) - \lambda_z(s) + \gamma \lambda_z(s')}{\alpha p(z)} \right) \right] \\ & - \sum_z \beta(z) \left[\mathbb{E}_{s \sim \bar{d}^B} \left[\frac{q_z(s)}{\bar{d}^B(s)} \right] - 1 \right]. \end{aligned} \tag{50}$$

This recovers Eq. equation 15.

Under deterministic dynamics, the sampled next state is uniquely determined by (s, a), and the sample-based objective coincides with the corresponding exact dual objective. Under stochastic dynamics, the replacement is generally biased because f^* is nonlinear. However, since f^* is convex, Jensen's inequality gives

$$\mathbb{E}_{s' \sim P(\cdot|s,a)} \left[f^*(X(s')) \right] \geq f^* \left(\mathbb{E}_{s' \sim P(\cdot|s,a)}[X(s')] \right),$$

so the sample-based term provides an upper bound on the term involving the exact conditional expectation.

Why skill annotations are not required. A useful consequence of Eq. equation 50 is that no skill-conditioned dataset distribution such as d_z^B or $\bar{d}_z^B$ appears in the objective. Transitions are sampled only from the skill-agnostic distribution $(s, a, s') \sim d^B(s, a) P(s' \mid s, a)$, while skills enter through the prior $p(z)$ and the learned functions q_z and λ_z. Equivalently, during optimization we may independently sample $(s, a, s') \sim d^B,\ z \sim p(z)$. Thus, the skill-conditioned target occupancies are created by the optimization itself rather than inherited from skill labels in the dataset. This is the key reason MIM-DICE can learn skills directly from arbitrary off-policy experience.

DDPG Hyperparameter	Value
Replay buffer capacity	10^6
Action repeat	1
Seed frames	4000
n-step returns	3
Mini-batch size	1024
Discount γ	0.99
Optimizer	Adam
Learning rate	1×10^{-4}
Agent update frequency	2
Critic target EMA rate τ_c	0.01
Feature dimension	1024
Hidden dimension	1024
Exploration stddev clip	0.3
Exploration stddev value	0.2
Initialize critic	False
Pre-training frames	MAZE: 1×10^6; URLB: 2×10^6
Fine-tuning frames	URLB: 1×10^5
MIM-DICE-specific Hyperparameter	**Value**
MIM coefficient α	0.05
f-divergence	Soft-χ^2
Hidden dimension	1024
Number of hidden layers	1
Skill dimension (URLB)	64
Skill prior (URLB)	$z \sim \mathrm{Unif}([0,1]^{64})$

Table 2: Hyperparameters used for the underlying DDPG agent and MIM-DICE. MIM-DICE uses the same optimizer, learning rate, batch size, discount factor, and n-step return settings as the underlying agent unless otherwise specified.

B Experimental Details

This section provides additional experimental and implementation details for the evaluations in Section 5. We first describe the common implementation choices and the practical optimization procedure for MIM-DICE, and then provide environment-specific details for the tabular, Pendulum, MAZE, and URLB experiments. Unless otherwise specified, we follow the standard URLB implementation and evaluation protocol.

B.1 Implementation Details

Network architecture. For the URLB experiments, MIM-DICE uses a 64-dimensional continuous skill variable

$$z \sim \mathrm{Unif}([0,1]^{64}). \tag{51}$$

All additional function approximators introduced by MIM-DICE use a single hidden layer with width 1024. The underlying actor and critic architectures follow the standard URLB implementation.

Density estimation. The MIM-DICE objective requires an estimate of the replay-buffer state density $\bar{d}^B(s)$. We use the same state-density estimation procedure and associated hyperparameters as APT (Liu & Abbeel, 2021b) and SEMDICE (Lee et al., 2025). We do not introduce additional tuning of the density estimator for MIM-DICE.

Optimization. MIM-DICE optimizes the saddle-point objective in Eq. equation 15 by alternating between the dual variables (λ, μ, β) and the auxiliary distributions q. For each minibatch, skill variables are sampled independently from $p(z)$ on-the-fly and paired with transitions sampled from the replay buffer. The dual variables are updated by gradient descent while holding q fixed, and q is updated by gradient ascent while holding the dual variables fixed. The resulting procedure implements the approximate MM optimization described in Section 3.

Algorithm 1 Alternating Optimization of MIM-DICE

Require: Replay buffer $\mathcal{B}$, skill prior $p(z)$
Require: Dual variables (λ, μ, β) and auxiliary model q
1: **repeat**
2: Sample a transition minibatch $\{(s_i, a_i, s'_i)\}$ from $\mathcal{B}$.
3: Sample skill codes $\{z_k\}_{k=1}^{K} \sim p(z)$ on-the-fly.
4: Form state–skill pairs (s_i, z_k) and (s'_i, z_k).
5: Evaluate $\mu(s_i)$, $\lambda_{z_k}(s_i)$, $\lambda_{z_k}(s'_i)$, $\beta(z_k)$, and $\log q_{z_k}(s_i)$.
6: Update (λ, μ, β) by gradient descent:
$(\lambda, \mu, \beta) \leftarrow (\lambda, \mu, \beta) - \eta_{\text{dual}} \nabla_{\lambda,\mu,\beta} \mathcal{L}_{\text{MIM}}\big(\lambda, \mu, \beta, \text{sg}(q)\big)$.
7: Update q by gradient ascent:
$q \leftarrow q + \eta_q \nabla_q \mathcal{L}_{\text{MIM}}\big(\text{sg}(\lambda), \text{sg}(\mu), \text{sg}(\beta), q\big)$.
8: Update the skill-conditioned policy using the information projection described in Appendix C.1.
9: **until** convergence

B.2 Offline MAZE Experiments

Environment and baseline setup. We use the MAZE environments and dataset construction from the official ExDM implementation without modification. All methods are trained only on the resulting fixed offline datasets.

To the best of our knowledge, prior skill-discovery methods do not directly target fully offline learning with continuous skill variables. We therefore compare against APS and CIC, which support continuous skills and optimize $H(S) - H(S \mid Z)$-based objectives. For both methods, we use the implementations provided by URLB without method-specific modifications.

C Practical Implementation

This section describes how the sample-based MIM-DICE objective is optimized in practice and how the learned occupancy correction is used to extract a skill-conditioned policy. We first describe the policy-extraction procedure and then provide the practical optimization details used in our experiments.

MIM-DICE specifically addresses the replay–policy distribution mismatch that arises in mutual-information optimization. In practice, however, its performance may still be affected by standard approximation errors associated with function approximation, including inexact MM updates, and policy extraction. Code for reproducing our experiments is available at: `https://anonymous.4open.science/r/mimdice-anonymous-4ED5/README.md`.

C.1 Policy Extraction

The optimization in Eq. equation 15 produces the dual variables and auxiliary distribution that characterize the target skill-conditioned occupancies, but does not directly produce an explicit policy. We therefore recover the target occupancy ratio from the optimal dual solution and use it to train a skill-conditioned policy through an information projection.

Recovering the occupancy ratio. Recall from the dual derivation that

$$f^*(y) = \sup_{w \geq 0} \{wy - f(w)\}. \tag{52}$$

For the optimal dual variables $(\lambda^\star, \mu^\star)$ and auxiliary distribution $q^\star$, define

$$e_{\lambda^\star,\mu^\star,q^\star}(s,a,z) := p(z)\left(\log q_z^\star(s) + \mu^\star(s)\right) - \lambda_z^\star(s) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\lambda_z^\star(s')\right].$$

The corresponding optimal density ratio

$$w_z^\star(s,a) := \frac{d_z^\star(s,a)}{d^B(s,a)} \tag{53}$$

satisfies

$$w_z^\star(s,a) \in \arg\max_{w\geq 0} \left\{ w \frac{e_{\lambda^\star,\mu^\star,q^\star}(s,a,z)}{\alpha p(z)} - f(w) \right\}. \tag{54}$$

When f^* is differentiable, this reduces to

$$w_z^\star(s,a) = \nabla f^* \left(\frac{e_{\lambda^\star,\mu^\star,q^\star}(s,a,z)}{\alpha p(z)} \right). \tag{55}$$

Thus, the optimal dual solution provides the occupancy correction required to recover the target skill-conditioned behavior. The remaining step is to extract a policy that realizes this corrected occupancy.

Information projection. Let $\pi_D(a \mid s)$ denote the behavior policy induced by the off-policy data distribution,

$$\pi_D(a \mid s) := \frac{d^B(s,a)}{\bar{d}^B(s)}. \tag{56}$$

Using $d_z^\star(s,a) = w_z^\star(s,a) d^B(s,a)$, the policy associated with the target occupancy satisfies

$$\pi_z^\star(a \mid s) = \frac{d_z^\star(s,a)}{\sum_{\tilde{a}} d_z^\star(s,\tilde{a})} = \frac{w_z^\star(s,a)\pi_D(a \mid s)}{\sum_{\tilde{a}} w_z^\star(s,\tilde{a})\pi_D(\tilde{a} \mid s)}.$$

We extract the skill-conditioned policy by minimizing the I-projection

$$\pi^\star \in \arg\min_\pi \; \mathbb{E}_{\substack{s\sim\bar{d}^B \\ z\sim p(z)}} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid s,z) \parallel \pi_z^\star(\cdot \mid s) \right) \right]. \tag{57}$$

Substituting Eq. equation **??** gives

$$\begin{aligned} &D_{\mathrm{KL}} \left(\pi(\cdot \mid s,z) \parallel \pi_z^\star(\cdot \mid s) \right) \\ &= \mathbb{E}_{a\sim\pi(\cdot|s,z)} \left[-\log w_z^\star(s,a) + \log \frac{\pi(a \mid s,z)}{\pi_D(a \mid s)} \right] + C(s,z), \end{aligned} \tag{58}$$

where

$$C(s,z) = \log \mathbb{E}_{a\sim\pi_D(\cdot|s)} \left[w_z^\star(s,a) \right] \tag{59}$$

is independent of π. Therefore, policy extraction is equivalent to

$$\min_\pi \quad \mathbb{E}_{\substack{s\sim\bar{d}^B,\, z\sim p(z) \\ a\sim\pi(\cdot|s,z)}} \left[-\log w_z^\star(s,a) + \log \frac{\pi(a \mid s,z)}{\pi_D(a \mid s)} \right]. \tag{60}$$

Equivalently,

$$\min_\pi \quad \mathbb{E}_{\substack{s\sim\bar{d}^B,\, z\sim p(z) \\ a\sim\pi(\cdot|s,z)}} \left[-\log w_z^\star(s,a) \right] + \mathbb{E}_{s\sim\bar{d}^B,\, z\sim p(z)} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid s,z) \parallel \pi_D(\cdot \mid s) \right) \right]. \tag{61}$$

This recovers the policy-extraction objective used in the section 3. The first term encourages the policy to select actions with high target occupancy ratio, while the KL term regularizes the extracted policy toward actions supported by the off-policy dataset.

Learning the occupancy-ratio model. Direct evaluation of $e_{\lambda,\mu,q}(s,a,z)$ requires the conditional expectation $\mathbb{E}_{s'\sim P(\cdot|s,a)}[\lambda_z(s')]$. For policy extraction, we therefore learn a residual model $e_\xi(s,a,z)$ from replay-buffer transitions.

For a sampled transition $(s,a,s') \sim \mathcal{B}$, define the single-sample residual

$$\widehat{e}_{\lambda,\mu,q}(s,a,s',z) := p(z) \left(\log q_z(s) + \mu(s) \right) - \lambda_z(s) + \gamma \lambda_z(s'). \tag{62}$$

We fit e_ξ by minimizing

$$\min_\xi \; \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{B} \\ z\sim p(z)}} \left[\left(e_\xi(s,a,z) - \widehat{e}_{\lambda,\mu,q}(s,a,s',z) \right)^2 \right]. \tag{63}$$

The corresponding occupancy-ratio estimate is

$$w_z^{\xi}(s,a) = \nabla f^* \left(\frac{e_{\xi}(s,a,z)}{\alpha p(z)} \right). \tag{64}$$

In practice, w_z^{ξ} is substituted for $w_z^{\star}$ in Eq. equation 61 when updating the skill-conditioned actor.

With a parameterized skill-conditioned policy $\pi_\theta(a \mid s, z)$, we use the following practical policy-extraction objective:

$$\theta^{\star} \in \arg\min_{\theta} \mathbb{E}_{s\sim\bar{d}^B,\, z\sim p(z),\, a\sim\pi_\theta(\cdot|s,z)} \left[-\log w_z^{\xi}(s,a) + \eta_{\mathrm{BC}} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s,z) \,\|\, \pi_B(\cdot \mid s)) \right], \tag{65}$$

where $\eta_{\mathrm{BC}} \geq 0$ controls the contribution of the behavior-cloning term. Setting $\eta_{\mathrm{BC}} = 1$ corresponds to the I-projection objective, whereas $\eta_{\mathrm{BC}} = 0$ reduces to policy extraction using only the recovered occupancy ratio.

In practice, estimating the behavior policy π_B from arbitrary off-policy data can introduce additional approximation error and sensitivity. We therefore fix $\eta_{\mathrm{BC}} = 0$ for all URLB domains, rather than tuning it separately for each domain.

C.2 Downstream Skill Selection

During downstream adaptation, MIM-DICE must select a skill whose pretrained behavior is aligned with the newly observed extrinsic reward. Rather than enumerating skills, which is infeasible for the continuous skill space used in URLB, we optimize the skill variable directly using a reward-basis regression objective.

Recall that the Bellman-like residual learned by MIM-DICE contains the state–skill dependent term

$$\mu(s) + \log q_z(s). \tag{66}$$

Motivated by its role as the reward-like component of the OptiDICE-style residual (Lee et al., 2021), we define the MIM-DICE reward basis as $\Psi_{\mathrm{MIM}}(s,z) := \mu(s) + \log q_z(s)$. The pretrained functions μ and q are kept fixed during the skill-selection procedure.

Given a fine-tuning replay buffer $\mathcal{B}_{\mathrm{ft}}$, we select a skill by fitting the observed downstream reward with an affine transformation of the reward basis:

$$\min_{z,\kappa,b} \quad \mathcal{L}_{\mathrm{select}}(z,\kappa,b) = \mathbb{E}_{(s,r_{\mathrm{ext}})\sim\mathcal{B}_{\mathrm{ft}}} \left[\left(\kappa\, \Psi_{\mathrm{MIM}}(s,z) + b - r_{\mathrm{ext}}(s) \right)^2 \right], \tag{67}$$

where $\kappa > 0$ and $b \in \mathbb{R}$ account for task-dependent scale and shift in the downstream reward. Because z is continuous, the regression can be optimized directly with respect to the skill variable.

After obtaining $z^{\star} = \arg\min_z \mathcal{L}_{\mathrm{select}}(z,\kappa,b)$, we fix $z^{\star}$ and condition the pretrained policy on this skill during downstream adaptation. The policy and critic are then updated using the extrinsic reward r_{ext} with the standard off-policy RL procedure.

In our implementation, skill selection is performed once after the fine-tuning buffer contains initial task-relevant experience. We optionally repeat the regression once using the updated fine-tuning buffer. All other downstream policy and critic updates follow the standard URLB protocol.

D Additional Experimental Results

D.1 Per-Task URLB Results

Table 3 reports the expert-normalized downstream performance on each URLB task. The aggregate results reported in the main text are computed from these per-task scores using the RLiable evaluation protocol.

Algorithm 2 Downstream Skill Selection via Reward-Basis Regression

Require: Fine-tuning buffer $\mathcal{B}_{\text{ft}}$
Require: Frozen pretrained functions $\mu(s)$ and $q_z(s)$
Require: Skill prior $p(z)$, optimization iterations M
Require: Selection step T_{select}
Initialize $z^\star \sim p(z)$
for each fine-tuning step $t = 1, 2, \ldots$ **do**
  if $t = T_{\text{select}}$ **then**
    Initialize $z \sim p(z)$, $\kappa \leftarrow 1$, and $b \leftarrow 0$
    for $i = 1, \ldots, M$ **do**
      Sample $\{(s, r_{\text{ext}})\} \sim \mathcal{B}_{\text{ft}}$
      Compute

$$\mathcal{L}_{\text{select}} = \mathbb{E}_{\mathcal{B}_{\text{ft}}}\left[\left(\kappa\big(\mu(s) + \log q_z(s)\big) + b - r_{\text{ext}}(s)\right)^2\right]$$

      Update (z, κ, b) by gradient descent
    end for
    $z^\star \leftarrow z$
  end if
  Execute $a_t \sim \pi(\cdot \mid s_t, z^\star)$
  Update the policy and critic using r_{ext}
end for

Table 3: Expert-normalized scores across downstream tasks. Results are reported as mean ± SE.

Domain	Task	MIM-DICE	DOI [16]	ExDM	ICM	Disag.	RND	APT	Proto.	CESD	BECL	CIC	SMM	DIAYN	APS
Walker	Flip	70.3 ± 2.6	56.1 ± 3.6	71.2 ± 2.9	60.1 ± 1.2	62.8 ± 3.2	61.8 ± 1.9	64.5 ± 2.1	53.3 ± 1.6	57.2 ± 3.2	73.0 ± 2.4	68.6 ± 2.7	68.7 ± 2.6	46.2 ± 2.5	69.0 ± 3.6
	Run	43.1 ± 2.6	25.7 ± 4.0	44.8 ± 3.7	40.5 ± 4.9	41.3 ± 3.8	45.6 ± 3.8	42.0 ± 2.7	42.0 ± 4.5	23.1 ± 2.9	34.9 ± 3.0	40.2 ± 2.7	45.6 ± 4.4	21.5 ± 0.9	39.1 ± 2.3
	Stand	95.8 ± 0.6	85.7 ± 4.3	97.1 ± 0.4	94.2 ± 1.3	92.3 ± 1.5	92.7 ± 0.6	92.4 ± 0.9	92.4 ± 0.7	92.7 ± 2.1	93.1 ± 1.5	93.1 ± 2.6	95.8 ± 0.7	75.2 ± 2.3	83.5 ± 8.3
	Walk	90.3 ± 3.4	74.0 ± 6.5	95.0 ± 2.0	84.5 ± 4.3	88.7 ± 2.5	92.5 ± 3.4	89.8 ± 2.8	78.6 ± 1.5	74.6 ± 4.5	93.1 ± 2.9	84.7 ± 3.4	94.0 ± 2.1	50.7 ± 4.4	78.7 ± 5.6
Quadruped	Jump	49.9 ± 6.7	35.8 ± 6.8	67.5 ± 4.5	30.5 ± 3.8	38.5 ± 5.3	49.8 ± 8.1	59.1 ± 5.0	36.7 ± 5.3	69.7 ± 4.1	36.6 ± 3.7	64.4 ± 6.0	46.4 ± 5.6	32.3 ± 5.5	34.0 ± 6.3
	Run	29.7 ± 4.1	24.8 ± 3.8	43.0 ± 1.7	18.6 ± 2.5	30.8 ± 5.9	31.7 ± 3.3	38.0 ± 3.6	22.0 ± 2.8	46.3 ± 1.5	23.4 ± 3.5	33.8 ± 3.5	27.5 ± 4.5	21.9 ± 3.4	23.2 ± 2.0
	Stand	68.6 ± 7.5	45.1 ± 7.0	89.2 ± 3.5	54.0 ± 3.2	48.3 ± 6.3	75.5 ± 8.2	64.2 ± 8.9	53.9 ± 7.7	70.9 ± 9.8	48.9 ± 3.9	69.5 ± 6.8	56.0 ± 6.3	52.2 ± 8.0	49.4 ± 5.7
	Walk	26.9 ± 5.4	21.3 ± 3.4	69.7 ± 6.1	17.7 ± 2.4	36.8 ± 6.1	34.1 ± 3.5	42.2 ± 6.2	25.6 ± 3.2	56.2 ± 6.8	24.3 ± 5.2	34.1 ± 9.5	28.0 ± 4.1	21.8 ± 3.1	19.1 ± 2.8
Jaco	Reach BL	82.9 ± 5.9	54.6 ± 10.2	32.1 ± 9.7	38.3 ± 4.8	22.8 ± 7.1	23.8 ± 9.5	17.1 ± 7.8	22.8 ± 9.0	50.3 ± 6.1	37.8 ± 7.6	79.8 ± 5.5	32.6 ± 4.8	1.0 ± 1.0	41.5 ± 13.4
	Reach BR	76.7 ± 8.4	54.3 ± 8.8	45.7 ± 11.8	28.6 ± 7.9	24.6 ± 7.9	20.6 ± 5.9	18.6 ± 7.8	20.6 ± 7.9	50.3 ± 9.1	62.8 ± 6.7	74.4 ± 5.0	30.2 ± 6.1	0.0 ± 0.0	57.8 ± 10.0
	Reach TL	79.6 ± 9.1	51.4 ± 7.4	47.1 ± 9.2	30.4 ± 12.6	24.6 ± 5.3	13.6 ± 3.6	18.8 ± 5.6	13.6 ± 9.1	22.0 ± 8.5	35.6 ± 8.3	69.6 ± 7.6	25.1 ± 8.0	1.0 ± 0.5	45.0 ± 11.5
	Reach TR	84.6 ± 5.6	53.0 ± 9.6	23.2 ± 12.7	21.2 ± 7.4	31.5 ± 11.4	19.2 ± 6.6	7.4 ± 2.8	17.7 ± 6.1	23.2 ± 7.5	25.6 ± 5.2	78.8 ± 4.6	23.6 ± 9.5	0.0 ± 0.0	38.4 ± 8.5

D.2 Additional Comparison with DOI

We additionally evaluate DOI (Vlastelica et al., 2024) under the same Pendulum setting used in the main experiments. Table 4a extends the comparison in Table 1a with the DOI result. DOI achieves a policy MI of only 0.038, indicating that the learned skills are nearly indistinguishable. Further inspection shows that DOI's skill discriminator collapses toward the uniform distribution. We observe the same behavior in the URLB domains, as shown in Table 4b. DOI (Vlastelica et al., 2024) can still retain non-trivial downstream performance 3, partly due to its auxiliary information-gain signal, despite the eventual discriminator collapse.

This behavior is consistent with the discriminator-collapse failure mode reported in the original DOI (Vlastelica et al., 2024) paper. Although DOI and MIM-DICE both use DICE-style occupancy optimization, they optimize different MI formulations: DOI is based on $H(Z) - H(Z \mid S)$. In our experiments, MIM-DICE does not exhibit the uniform-skill collapse observed with DOI.

D.3 Variational Gap During Optimization

The MM formulation introduces the auxiliary distribution q_z to majorize the conditional-entropy term. To examine how tightly this auxiliary distribution tracks the learned skill-conditioned occupancy, we measure the corresponding variational gap during training. For this diagnostic, we extend the tabular training run to 5M steps.

Table 4c shows that the gap decreases substantially over training, from 0.714 ± 0.026 at the beginning to 0.088 ± 0.030 in the final interval. This indicates that the auxiliary distribution increasingly tracks the optimized skill-conditioned occupancy as training progresses.

Table 4: Additional diagnostics and comparisons. (a): Pendulum performance including DOI. (b): DOI discriminator collapse across domains. (c): variational gap during MIM-DICE optimization.

Method	$\widehat{I}(S;Z)$ (↑)
CIC	0.13 ± 0.05
SMM	0.17 ± 0.06
APS	0.60 ± 0.12
DIAYN	0.66 ± 0.08
BECL	0.37 ± 0.09
CESD	0.02 ± 0.01
DOI	0.038 ± 0.01
MIM-DICE (Ours)	$\mathbf{0.94 \pm 0.05}$

(a) Pendulum performance

Domain	$D_{\mathrm{KL}}(q \,\|\, U)$
Pendulum	0.0001 ± 0.0001
Walker	0.0001 ± 0.0001
Quadruped	0.0001 ± 0.0001
Jaco	0.0002 ± 0.0000

(b) DOI discriminator collapse

Step Interval	**Variational Gap** (↓)
0–200K	0.714 ± 0.026
200–400K	0.820 ± 0.092
400–600K	0.344 ± 0.077
600–800K	0.165 ± 0.049
800K–1M	0.139 ± 0.070
1–2M	0.120 ± 0.054
2–3M	0.092 ± 0.032
3–4M	0.089 ± 0.031
4–5M	0.088 ± 0.030

(c) Variational gap

Table 5: Sensitivity analysis of MIM-DICE on Pendulum. Left: sensitivity to the choice of f-divergence. Right: sensitivity to the regularization coefficient α. Results are reported as mean ± SE.

f**-divergence**	$\widehat{I}(S;Z)$	$\widehat{H}(S)$	$\widehat{H}(S \mid Z)$
KL	0.581 ± 0.156	3.622 ± 0.258	3.041 ± 0.165
Pearson χ^2	$\mathbf{0.957 \pm 0.033}$	3.869 ± 0.078	2.912 ± 0.059
SoftChi	0.943 ± 0.048	3.731 ± 0.122	2.789 ± 0.129

(a) f-divergence sensitivity

α	$\widehat{I}(S;Z)$	$\widehat{H}(S)$	$\widehat{H}(S \mid Z)$
0.1	$\mathbf{0.943 \pm 0.048}$	3.731 ± 0.122	2.789 ± 0.129
0.2	0.938 ± 0.071	3.605 ± 0.052	2.667 ± 0.073
0.5	0.903 ± 0.070	3.527 ± 0.103	2.624 ± 0.077
1.0	0.352 ± 0.112	3.456 ± 0.072	3.104 ± 0.183

(b) α sensitivity with SoftChi

D.4 Hyperparameter Sensitivity

We evaluate the sensitivity of MIM-DICE to the choice of f-divergence and the regularization coefficient α using the Pendulum setting from the main text. The results are summarized in Table 5.

For the choice of f-divergence, Pearson χ^2 and SoftChi achieve similarly high MI, whereas KL yields lower performance with substantially higher variance. This suggests that Pearson χ^2 and SoftChi provide more numerically stable optimization in this setting.

For $\alpha \in \{0.1, 0.2, 0.5\}$, the learned policies consistently achieve high $\widehat{I}(S;Z)$. In contrast, performance degrades at $\alpha = 1.0$. In particular, $\widehat{H}(S \mid Z)$ increases from 2.624 at $\alpha = 0.5$ to 3.104 at $\alpha = 1.0$. This behavior is consistent with overly strong f-divergence regularization constraining the skill-conditioned occupancies to remain close to the replay-buffer distribution, thereby limiting their separation.

Overall, our results suggest using Pearson χ^2 or SoftChi together with a moderately small α (e.g., $\alpha \leq 0.5$) in practice.

E Applying MIM-DICE to Stochastic Dynamics

As discussed in the section 3, under stochastic dynamics, replacing the conditional expectation inside the nonlinear function f^* with a single transition sample generally introduces a Jensen bias. MIM-DICE can be extended to stochastic environments using a semi-gradient update.

For each z, define $e_z(s,a,s') = p(z)\big(\log q_z(s) + \mu(s)\big) - \lambda_z(s) + \gamma\lambda_z(s')$. The quantity required inside f^* under stochastic dynamics is $\bar{e}_z(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[e_z(s,a,s')\right]$.

We estimate this conditional expectation using an auxiliary network $e_{z,\phi}(s,a)$ by minimizing

$$\min_{\phi} \; \frac{1}{2}\mathbb{E}_{(s,a,s') \sim d^{\mathcal{B}}}\left[e_{z,\phi}(s,a) - \mathrm{sg}(e_z(s,a,s'))\right]^2 .$$

Then, instead of evaluating $(f^*)'$ at the single-sample quantity $e_z(s, a, s')$, we evaluate it at the estimated conditional expectation and use the following semi-gradient surrogate:

$$\mathbb{E}_{(s,a,s')\sim d^{\mathcal{B}}}\left[\mathrm{sg}\left[(f^*)'\left(\frac{e_{z,\phi}(s,a)}{\alpha p(z)}\right)\right] e_z(s,a,s')\right].$$

At the population optimum, $e_{z,\phi}(s, a) = \bar{e}_z(s, a)$, so the nonlinear weight $(f^*)'$ is evaluated at the correct conditional expectation rather than at a single stochastic transition. This avoids the Jensen bias arising from directly substituting a single next-state sample inside f^*.

F VISUALIZATION OF SIX MAZE DOMAINS

Coverage 50%

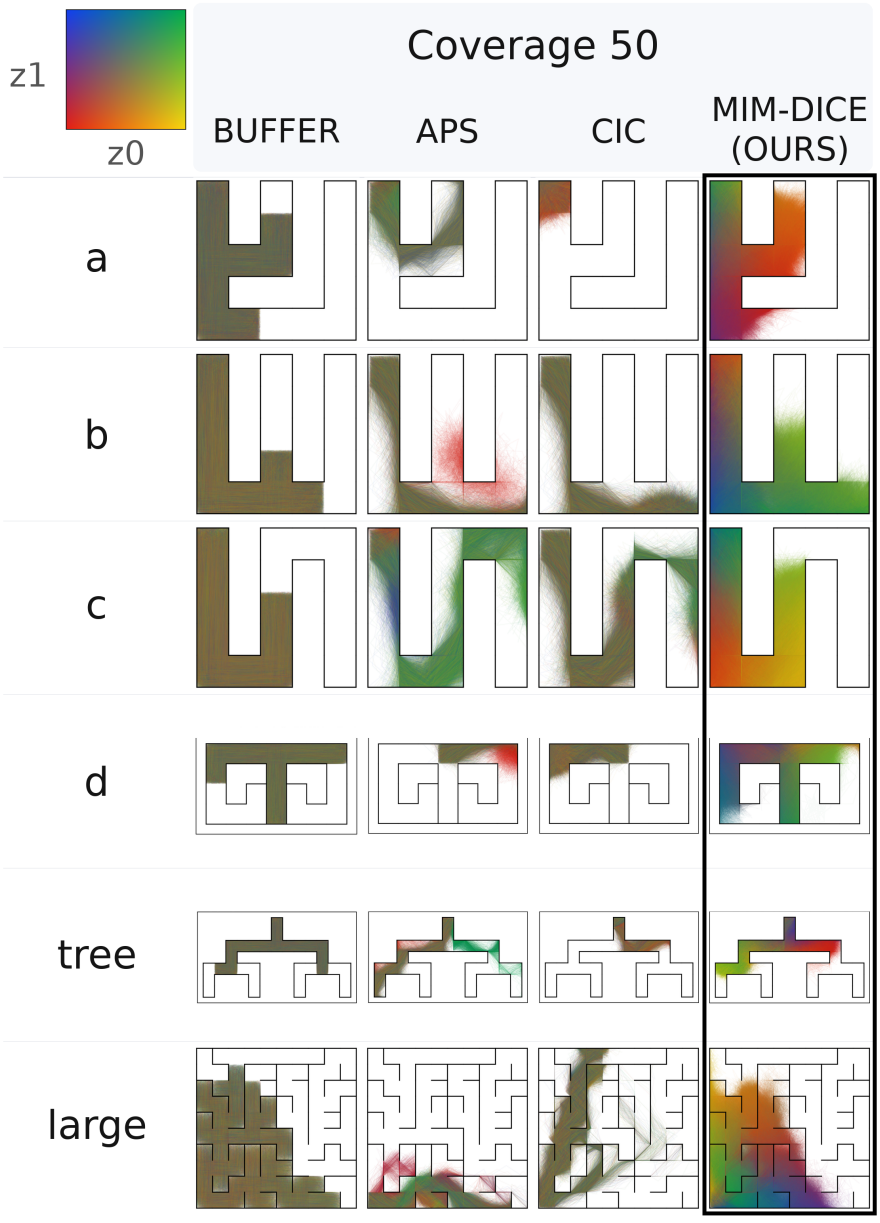


Figure 4: MAZE Buffer coverage 50%

Coverage 70%

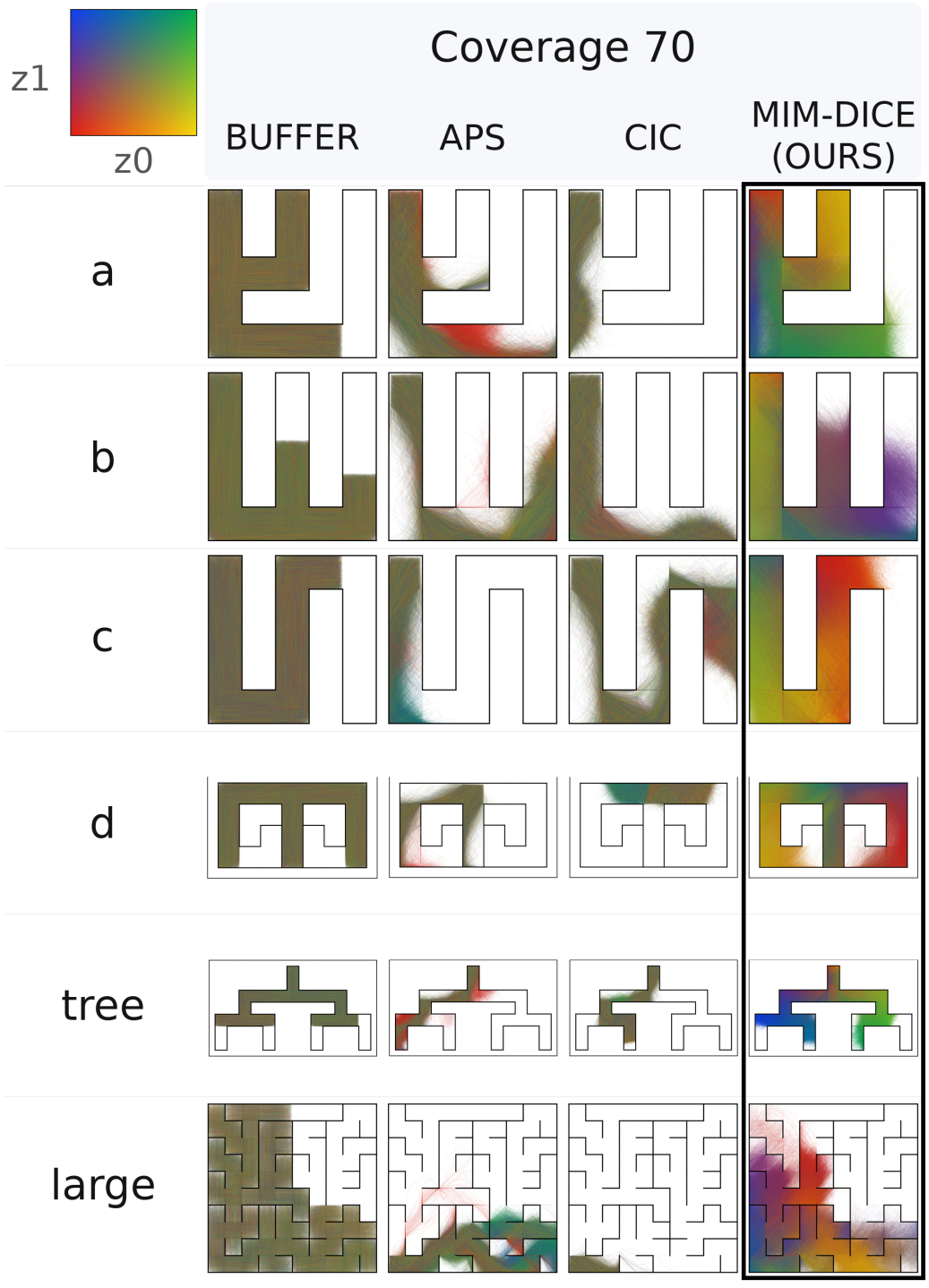


Figure 5: MAZE Buffer coverage 70%

Coverage 100%

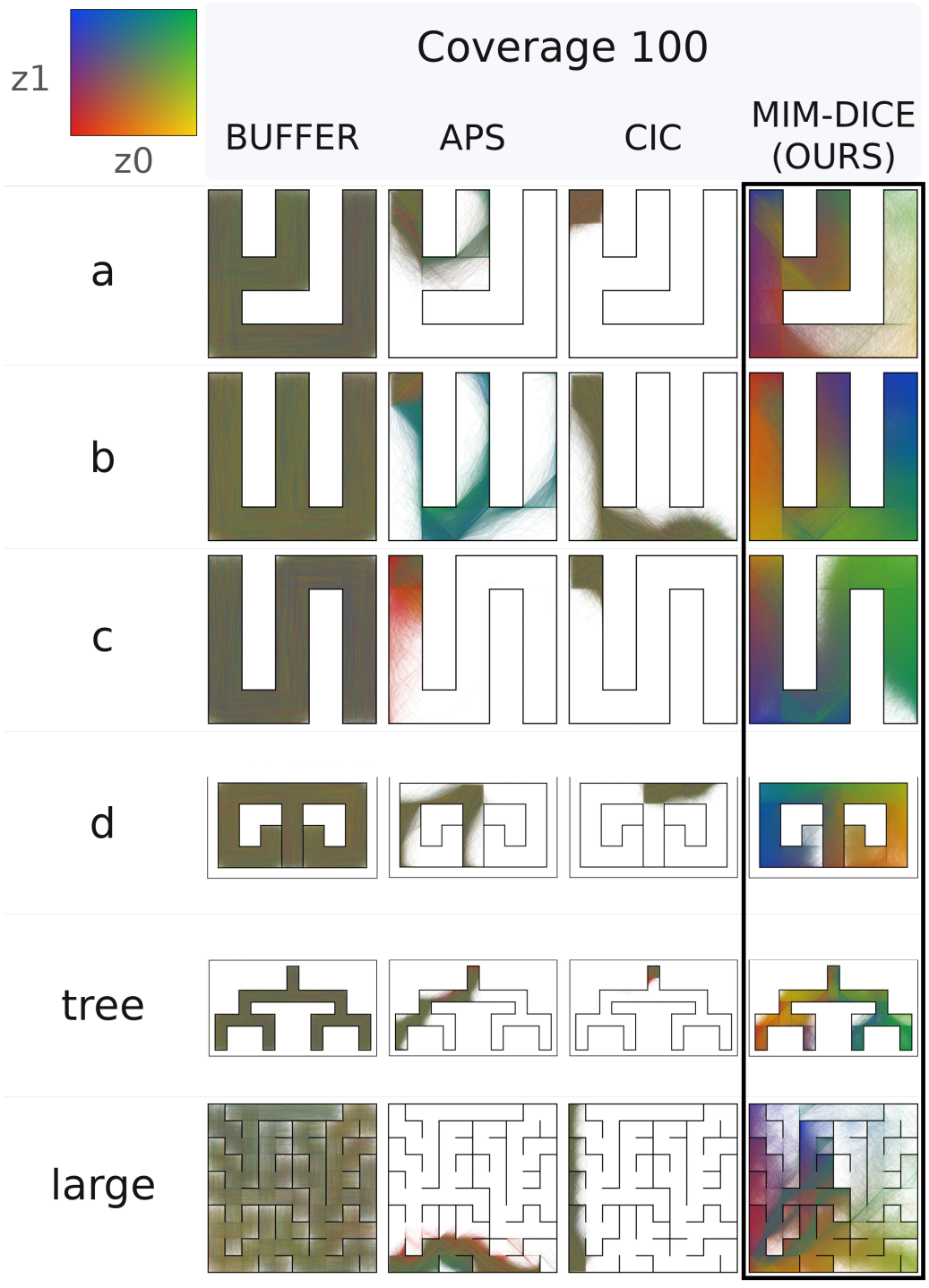


Figure 6: MAZE Buffer coverage 100%