Spectral Alignment in Forward–Backward Representations via Temporal Abstraction

Seyed Mahdi B. Azad Jasper Hoffmann Iman Nematollahi Hao Zhu

Abhinav Valada Joschka Bödecker

Department of Computer Science, University of Freiburg, Germany
basiri,hoffmaja,nematoli,zhuh,valada,jboedeck@cs.uni-freiburg.de

Abstract

Forward-backward (FB) representations provide a powerful framework for learning the successor representation (SR) in continuous spaces by enforcing a low-rank factorization. However, a fundamental spectral mismatch often exists between the high-rank transition dynamics of continuous environments and the low-rank bottleneck of the FB architecture, making accurate low-rank representation learning difficult. In this work, we analyze temporal abstraction as a mechanism to mitigate this mismatch. By characterizing the spectral properties of the transition operator, we show that temporal abstraction acts analogously to a low-pass filter that suppresses high-frequency spectral components. This suppression reduces the effective rank of the induced SR while preserving a formal bound on the resulting value function error. Empirically, we show that this alignment is a key factor for stable FB learning, particularly at high discount factors where bootstrapping becomes error-prone. Our results identify temporal abstraction as a principled mechanism for shaping the spectral structure of the underlying MDP and enabling effective long-horizon representations in continuous control.

1 Introduction

Effective long-horizon control requires representations that map current actions to future outcomes. The successor representation (SR) achieves this by encoding discounted future state–action occupancies (Dayan, 1993), providing a structured foundation for value computation across diverse rewards. While SR-based methods have successfully scaled to high-dimensional control (Kulkarni et al., 2016; Zhang et al., 2017), continuous domains require representations that are both expressive and computationally tractable. Forward-backward (FB) representations address this by learning a low-rank factorization of the SR directly from interaction (Blier et al., 2021; Touati & Ollivier, 2021).

However, a fundamental incompatibility exists: while FB assumes a low-rank constraint, the true SR in continuous environments is often high-rank with slow spectral decay (Dubail et al., 2025). We identify this spectral mismatch as a primary bottleneck for FB. Empirically, we show that increasing network capacity does not reliably improve performance; instead, higher capacity can lead to performance degradation as networks attempt to resolve high-frequency dynamical components that are inherently difficult to predict. When coupled with bootstrapping, errors in these spectral components propagate through Bellman updates and destabilize the learning process.

To address this, we leverage temporal abstraction via action repetition to regulate the SR's spectral structure. We demonstrate that multi-step transitions accelerate spectral decay, yielding a more structured, low-rank target for the FB objective. As shown in Figure 1, action repetition—previously

Preprint.

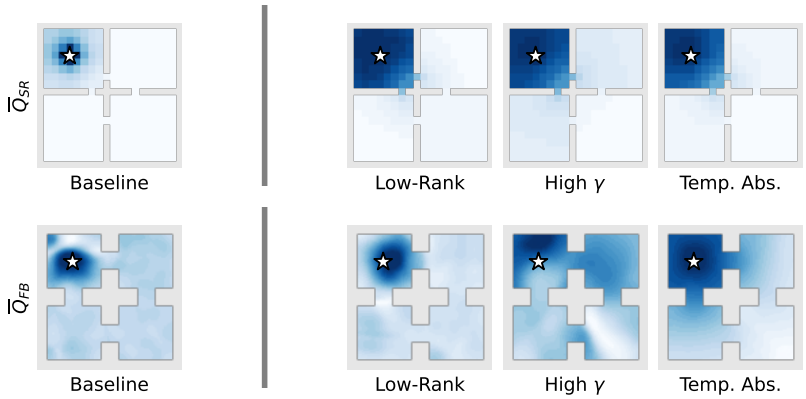


Figure 1: **Q-function via Successor Representation (SR).** The SR enables rapid value inference for arbitrary goals (e.g., *star marker*). Low-rank structure in SR is desirable for navigation, as it preserves topological features (e.g., rooms) while suppressing transient dynamics. **Top:** In discrete MDPs, the SR can be computed from the transition matrix. **Bottom:** In continuous domains, forward-backward (FB) learning approximates the SR, where the embedding dimension controls the rank of the approximation. Low-rank structure can arise through (1) explicit constraints (*Low-Rank* column; e.g., SVD or small embeddings), (2) long horizons (*High* γ column), or (3) temporal abstraction (*Temp. Abs.* column; e.g., action repetition). In continuous settings, temporal abstraction provides the spectral alignment needed for effective bootstrapping, whereas high γ or overly restrictive bottlenecks can impair representation learning, leading to Q-functions with many erroneous local maxima.

utilized for exploration and efficiency (Mnih et al., 2015; Biedenkapp et al., 2021)—consistently improves both representation quality and episodic return across discrete and continuous environments.

Finally, we examine the influence of the discount factor, γ. While increasing γ extends the task horizon, it also degrades SR conditioning and amplifies sub-dominant spectral components, increasing sensitivity to noise. We show that temporal abstraction counteracts this by improving spectral concentration, enabling stable learning at high effective horizons. Together, our results provide a unified perspective on the spectral requirements of FB learning, shifting the burden of representation from the function approximator to the design of interaction dynamics.

2 Related Works

Successor representations were introduced as task-agnostic predictive representations that enable rapid adaptation to new reward functions (Dayan, 1993). Subsequent work has leveraged SR for transfer and zero-shot reinforcement learning (Barreto et al., 2017). However, exact computation scales poorly with state dimensionality, motivating low-rank and parametric approximations that capture dominant long-horizon dynamics.

Forward-backward representation learning methods (Blier et al., 2021; Touati & Ollivier, 2021; Touati et al., 2023) address this challenge by learning factorizations of the SR that emphasize shared future occupancies over fine-grained state distinctions. While effective, these approaches implicitly rely on a low-rank SR structure and offer limited theoretical insight into when such a structure arises. Our work complements FB by linking the effective rank of the SR to the spectral properties of the transition dynamics induced by the policy and environment, and by proposing mechanisms that promote this low-rank structure.

The transition operator of a Markov decision process is central to long-term behavior, mixing, and value estimation. Classical Markov chain theory relates the spectral gap of the transition matrix to convergence rates (Meyn & Tweedie, 2012). In reinforcement learning, spectral methods have informed representation learning and planning, including proto-value functions and Laplacian-based abstractions (Mahadevan, 2005; Machado et al., 2017a,b; Shehmar et al., 2026). However, prior work focuses on policy evaluation and transfer, without analyzing how transition spectra influence the rank, compressibility, or learnability of low-rank SR under function approximation.

Temporal abstraction has been widely studied through semi-Markov decision processes and options (Sutton et al., 1999). A simple instance is action repetition (frame skipping), used in Atari benchmarks (Mnih et al., 2015) and known to significantly affect learning (Machado et al., 2018; Biedenkapp et al., 2021). From an operator perspective, repeating actions replaces the one-step transition matrix with its k-step counterpart, smoothing the dynamics. Existing work primarily motivates this via efficiency or exploration, without examining its impact on spectral structure or low-rank predictive representations such as the SR.

Although the connection between multi-step transitions and SR spectra is established (Dayan, 1993; Machado et al., 2017b,a; Dubail et al., 2025), and FB methods are empirically successful (Touati & Ollivier, 2021; Touati et al., 2023), their interaction remains underexplored. We reinterpret temporal abstraction not as an exploration heuristic (Lakshminarayanan et al., 2017), but as a spectral alignment mechanism that bridges high-rank dynamics and the low-rank inductive bias of FB representations.

3 Background

We represent a finite, reward-free Markov decision process (MDP) as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively, $P(s' \mid s, a)$ is the transition probability from state s to s' given action a, and $\gamma \in (0, 1)$ is the discount factor (Sutton & Barto, 1998). Given a policy π, the policy-induced transition operator is defined as a matrix $P^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$, where $P^\pi(s', a' \mid s, a) = \mathbb{P}(s_{t+1} = s', a_{t+1} = a' \mid s_t = s, a_t = a, \pi)$. The matrix P^π is row-stochastic, i.e., $P^\pi \mathbf{1} = \mathbf{1}$. The (discounted) SR associated with P^π is defined as $M^\pi = (I - \gamma P^\pi)^{-1} = \sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$. In the following, we use the matrix M^π and its functional form interchangeably. We define $M^\pi(s, a, s', a')$ as the expected discounted occupancy of (s', a') given an initial state-action pair (s, a). In matrix notation, it corresponds to the entry of M^π indexed by row (s, a) and column (s', a').

3.1 Forward-Backward Representation

The FB representation is a parametric framework designed to approximate the SR for all optimal policies in an unsupervised way (Touati & Ollivier, 2021). Let $(\pi_z)_{z\in\mathbb{R}^d}$ be a family of policies parameterized by $z \in \mathbb{R}^d$, and define the embedding functions $F\colon \mathcal{S}\times\mathcal{A}\times\mathbb{R}^d \to \mathbb{R}^d$ and $B\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$. Learning an FB representation entails finding (F, B, π_z) such that:

$$\pi_z(s) \in \operatorname*{argmax}_a F(s, a, z)^\top z \quad \text{and} \quad F(s, a, z)^\top B(s', a') = M^{\pi_z}(s, a, s', a') \tag{1}$$

for all $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$ and $z \in \mathbb{R}^d$. In continuous action spaces, the argmax in Eq. (1) is intractable. Following standard practice (Touati & Ollivier, 2021), we introduce a learned actor $\pi_\theta\colon \mathcal{S} \times \mathbb{R}^d \to \mathcal{A}$ trained jointly with F and B to approximate the maximizer. Architectural details and learning rates are deferred to Appendix A. Further, Eq. (1) represents a fixed-point condition for the triplet (F, B, π_z) since F and B depend on π_z, and π_z is defined via F (Touati & Ollivier, 2021). Given a reward function $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define $z_R = B^\top r$. If the condition holds exactly, the optimal action-value function is recovered by $Q^\star(s, a) = F(s, a, z_R)^\top z_R$.

Given a FB representation (F, B), we define the approximate successor representation as $\hat{M}^z(s, a, s', a') = F(s, a, z)^\top B(s', a')$. The following theorem bounds the approximation error of the optimal action-value function $Q^\star$ by the approximation error in successor representation M^{π_z}:

Theorem 3.1 (Optimality Gap for FB Representations)**.** *Let $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be a reward function such that $z_R = B^\top r$. The approximation error of the optimal Q-function is bounded by:*

$$\left\| F(\,\cdot\,,\,\cdot\,,\, z_R)^\top z_R - Q^\star \right\|_\infty \leq \frac{2C_{\text{norm}}\, \|r\|_\infty}{(1-\gamma)} \|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2. \tag{2}$$

Here, C_{norm} is a constant arising from the choices of norms, equal to $\sqrt{|\mathcal{S}||\mathcal{A}|}$ in our finite case; see Appendix B for details. Further, $\|\cdot\|_\infty$ denotes the L_∞ or Chebyshev norm, which for a function $f\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $\|f\|_\infty = \sup_{(s,a)} |f(s, a)|$. The norm $\|\cdot\|_2$ denotes the L_2 or spectral norm of a matrix, defined as $\|M\|_2 = \sup_{x\neq 0} \|Mx\|_2/\|x\|_2$, which corresponds to the largest singular value of M. Note that Theorem 3.1 is a simplified version of the result in (Touati & Ollivier, 2021, Theorem 8) tailored to our spectral analysis setting.

3.2 Spectral Bound on Approximation Error

To understand the approximation capacity of the FB framework, we derive a lower bound on the approximation error appearing on the right-hand side of Eq. (2) based on the spectrum of $M^{\pi_{z_R}}$. Related to this is the work in Dubail et al. (2025), which performs a similar study with a focus on finite-sample analysis. In this work, we do not aim to derive the tightest possible bound, but rather to develop a simple theoretical framework that highlights the effect of temporal abstractions on the optimal approximation error. We leave a finite-sample analysis to future work.

Due to limited representational capacity when d is small, the FB criterion cannot generally be fulfilled exactly, even in the finite case. Furthermore, the FB representation must simultaneously reconstruct the successor representation and define a greedy policy, as shown in Eq. (1). By the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), the best rank-d approximation of $M^{\pi_{z_R}}$ is obtained via the truncated singular value decomposition (SVD), denoted by $M^\star$, which satisfies $\|M^\star - M^{\pi_{z_R}}\|_2 = \sigma_{d+1}(M^{\pi_{z_R}})$. Intuitively, $\sigma_{d+1}(M^{\pi_{z_R}})$ corresponds to the first discarded singular value. Motivated by this observation, we define the following:

Definition 3.1 (Forward-backward Realization Error)**.** *Given a reward function $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, and a FB representation (F, B), we define the* FB realization error *as the difference to the optimal rank d approximation: $\epsilon_{\text{real}}(r) \coloneqq \|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2 - \sigma_{d+1}(M^{\pi_{z_R}}) \geq 0$.*

Consequently, the optimality gap in Eq. (2) is governed by the decay of the representation's singular values,

$$\left\|F(\,\cdot\,,\,\cdot\,,\,z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2C_{\text{norm}}\,\|r\|_\infty}{(1-\gamma)}\left(\epsilon_{\text{real}}(r) + \sigma_{d+1}(M^{\pi_{z_R}})\right),$$

assuming that the error $\epsilon_{\text{real}}(r)$ stays bounded. This decomposition separates the FB realization error $\epsilon_{\text{real}}(r)$ from the spectral truncation error $\sigma_{d+1}(M^{\pi_{z_R}})$ determined by the singular values of the successor representation.

4 Temporal Abstraction in Forward-Backward Representations

Our goal is to demonstrate that temporal abstraction is beneficial for learning FB representations. To this end, we introduce a simple temporal abstraction, namely action repetition. Action repetition was introduced in Mnih et al. (2015) and has been shown to be beneficial for exploration and learning performance in model-free reinforcement learning (RL) (Biedenkapp et al., 2021).

4.1 Action Repetition for Temporal Abstraction

In the following, we first formally introduce the concept of action-repeat MDPs, provide the necessary assumptions for this work, and conclude by connecting these concepts to the FB representation.

Definition 4.1 (Action-Repeat MDP)**.** *Given a reward-free MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma)$, an action-repeat MDP $\widetilde{\mathcal{M}}$ with repeat factor $k \in \mathbb{N}$ is defined by the tuple $(\mathcal{S}, \mathcal{A}, \widetilde{P}, \gamma^k)$. The transition probability $\widetilde{P}(s'|s,a)$ represents the probability of reaching state s' after executing action a for k consecutive time steps in $\mathcal{M}$. Mathematically, this is the k-fold composition of the transition operator:*

$$\widetilde{P}(s'|s,a) = \sum_{(s_1,\ldots,s_{k-1})\in S^{k-1}} P(s'|s_{k-1},a)\cdots P(s_1|s,a).$$

Note that for $k = 1$ we define $\widetilde{P}(s'|s,a) = P(s'|s,a)$.

Given this definition, and following Section 3, we define the successor representation $\widetilde{M}^\pi$ and the optimal state-action function $\widetilde{Q}^\star$ accordingly. To measure the error that is introduced by the action repetition, we introduce the following definition:

Definition 4.2 (Action-Repeat Value Error)**.** *For a given repeat factor k, we define the* action-repeat value error *as the worst-case discrepancy between the optimal Q-value function of the original MDP $\mathcal{M}$ and that of the action-repeat MDP $\widetilde{\mathcal{M}}$ as $\epsilon_{\text{repeat}}(k) \coloneqq \|Q^\star - \widetilde{Q}^\star\|_\infty$.*

For the remainder of this paper, we assume the existence of a repeat factor k such that the resulting action-repeat value error $\epsilon_{\text{repeat}}(k)$ is negligibly small. Furthermore, all representations (F, B) and successor measures $\hat{M}$ are hereafter assumed to be trained on the action-repeat MDP $\widetilde{\mathcal{M}}$.

4.2 Action Repetition Reduces the Optimality Gap

In the following, we will highlight that repeating each action for k steps introduces a trade-off between the action-repeat error $\epsilon_{\text{repeat}}(k)$ and an accelerated spectral decay. We denote the action-repeat policy-induced transition matrix by $\widetilde{P}^{\pi} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ with entries $\widetilde{P}^{\pi}((s,a),(s',a')) = \widetilde{P}(s' \mid s,a)\,\pi(a' \mid s')$, and write $P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with entries $P_a(s,s') = P(s' \mid s,a)$ for the single-step state-transition matrix under a fixed action a. We first combine our prior definitions to bound the overall approximation error. To derive this specific bound, we require the joint transition dynamics $\widetilde{P}^{\pi}$ to be diagonalizable, as stated in Assumption B.1.

Lemma 4.1 (Spectral Bound of Optimality Gap for k-repeat FB Representations)**.** *Given a reward function $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, let $\widetilde{r}(s,a) \coloneqq \mathbb{E}_P[\sum_{t=0}^{k-1}\gamma^t r(s_t,a) \mid s_0 = s]$ be the corresponding k-step expected reward function, and define $z_{\widetilde{r}} \coloneqq B^{\top}\widetilde{r}$. Under Assumption B.1, let (F,B) be an FB representation with dimension d. Then the error in approximating the original optimal action-value function $Q^{\star}$ is bounded by:*

$$\left\|F(\,\cdot\,,\,\cdot\,,z_{\widetilde{r}})^{\top}z_{\widetilde{r}} - Q^{\star}\right\|_{\infty} \le \epsilon_{\text{repeat}}(k) + \frac{2C_{\text{norm}}\,\|r\|_{\infty}}{1-\gamma}\left(\widetilde{\epsilon}_{\text{real}}(\widetilde{r}) + \frac{C_{\text{SF}}}{1-\gamma^k|\lambda_{d+1}(\widetilde{P}^{\pi})|}\right).$$

Here $|\lambda_{d+1}(\widetilde{P}^{\pi})|$ denotes the $(d{+}1)$-th largest absolute eigenvalue of $\widetilde{P}^{\pi}$, and $C_{\text{SF}} > 0$ is a constant from the spectral truncation of the successor representation; since $\widetilde{P}^{\pi}$ is row-stochastic and $\gamma^k < 1$, the denominator is automatically positive. Lemma 4.1 captures the trade-off between repetition error ϵ_{repeat}, the FB realization error $\widetilde{\epsilon}_{\text{real}}$, and the spectral truncation controlled by $|\lambda_{d+1}(\widetilde{P}^{\pi})|$.

As a next step, we examine how $|\lambda_{d+1}(\widetilde{P}^{\pi})|$ changes depending on the number of repeats k. Specifically, we relate the spectrum of $\widetilde{P}^{\pi}$ to that of the per-action transition matrices P_a collected into a single block-diagonal matrix $\mathbf{P}_{\mathcal{A}} \coloneqq \operatorname{diag}(P_{a_1},\ldots,P_{a_{|\mathcal{A}|}})$. As stated in Assumption B.2, we require the individual action blocks P_a to be diagonalizable to cleanly bound the spectrum of matrix powers.

Lemma 4.2 (Eigenvalue Contraction under Action Repetition)**.** *Under Assumption B.2, the $(d{+}1)$-th largest absolute eigenvalue of $\widetilde{P}^{\pi}$ contracts exponentially in k:*

$$|\lambda_{d+1}(\widetilde{P}^{\pi})| \;\le\; C_{\text{rep}}\,|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^{k}.$$

This establishes that the spectral term in Lemma 4.1 contracts at exponential rate $|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^k$ whenever $|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})| < 1$, which requires $d \ge |\mathcal{A}|$ since $\mathbf{P}_{\mathcal{A}}$ has $|\mathcal{A}|$ unit eigenvalues. The constant $C_{\text{rep}} > 0$ is a worst-case bound that grows with $|\mathcal{S}|$ and $|\mathcal{A}|$. The constants could potentially be tightened to $C_{\text{rep}} = \sqrt{|\mathcal{A}|}$ and $C_{\text{SF}} = 1$ under significantly stronger structural assumptions such as orthogonality. Proofs as well as details on the constants and assumptions are provided in Appendix B.

In practice, we believe the spectral decay of $\widetilde{P}^{\pi}$ to be significantly faster than suggested by the worst-case bound derived here.

5 Temporal Abstraction in Practice

We empirically validate the spectral insights from previous sections by examining how temporal abstraction shapes the structure and learnability of FB representations. We introduce spectral metrics for the effective rank of the SR, describe the experimental setup, and analyze how temporal abstraction reshapes the SR spectrum and affects performance. Finally, we study its interaction with embedding dimension and discount factor, highlighting their joint role in the stability and effectiveness of FB representation learning.

5.1 Spectral Metrics for Representation Complexity

To quantify the structure of the SR and the effect of temporal abstraction, we use two complementary spectral metrics.

Stable Rank. Stable rank captures how much spectral energy is concentrated in dominant directions. It decreases when a few leading components dominate, making it a direct proxy for low-rank approximability.

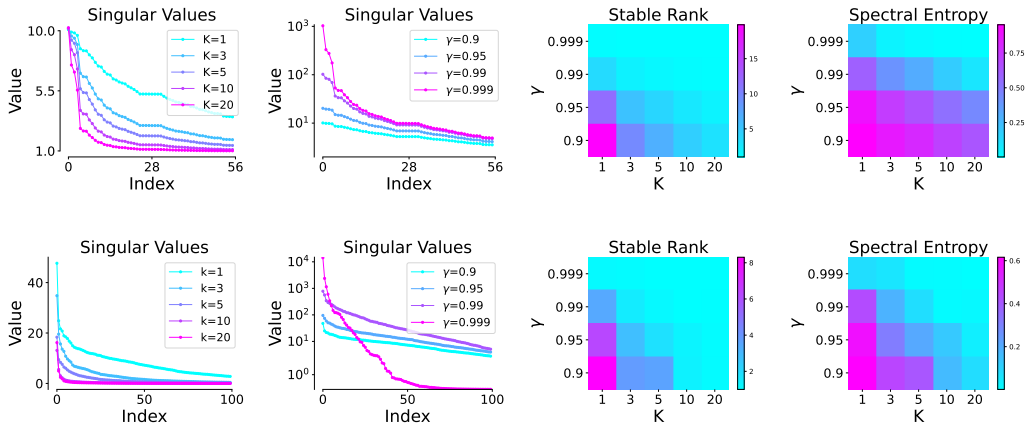


Figure 2: **Effect of temporal abstraction and discount factor on effective rank.** Effective rank decreases by increasing k or γ in both discrete *(top)* and continuous *(bottom)*. Entropy decreases more smoothly as k increases, suggesting a more stable reduction in effective rank as compared to increasing γ.

Normalized Spectral Entropy. Normalized spectral entropy measures how evenly spectral energy is distributed. High values indicate a diffused spectrum, while low values reflect concentration in a few components. Unlike stable rank, which emphasizes dominant modes, spectral entropy captures the overall spread of energy.

Together, these metrics characterize effective rank, distinguishing near rank-one collapse (low stable rank and entropy) from structured concentration, where few dominant components capture most energy while multiple modes remain active. Definitions of these metrics and details regarding their calculations for discrete and continuous settings are presented in Appendix C.

5.2 Experimental Setup

Notation. Throughout Sections 5 and 6 we follow practitioner usage: γ refers to the *nominal* discount used in training, i.e., the discount of the action-repeat MDP (corresponding to γ^k in the notation of Definition 4.1). The original-environment discount is then $\gamma_{\text{eff}} := \gamma^{1/k}$. We refer to γ_{eff} in Section 6 when this distinction matters for analyzing horizon trade-offs.

We evaluate temporal abstraction via action repetition in three continuous maze navigation tasks of increasing difficulty: *Four-Rooms*, *Maze*, and *Large-Maze* (Figure 3), implemented in OGBench (Park et al., 2025) with random start and goal positions.

Unless stated otherwise, experiments use Four-Rooms with discount factor $\gamma = 0.95$, embedding dimension $d = 100$, and action repetition $k = 10$. Following Touati & Ollivier (2021), states are encoded from (x, y) coordinates using an RBF kernel; similar results hold with learned CNN encoders (Figure 6a). We report the mean episodic return and the 95 percent confidence interval over five seeds. Each model is trained for one million gradient update steps. Key hyperparameters and implementation details are presented in Appendix A.

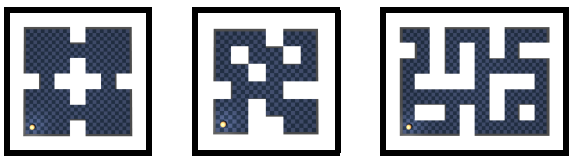
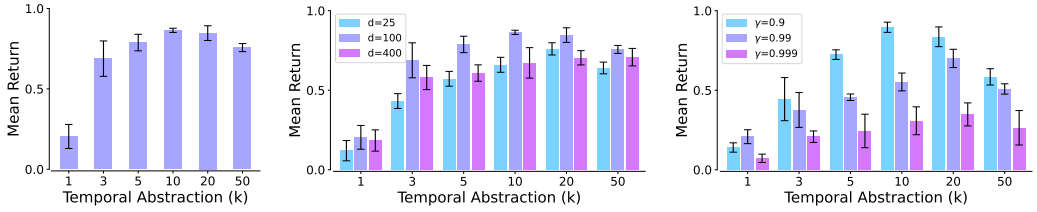


Figure 3: **Continuous Navigation Environments:** *Four-Rooms*, *Maze*, and *Large-Maze*



(a) (b) (c)

Figure 4: **Effect of temporal abstraction on performance.** Ablation over temporal abstraction (k), embedding dimension (d), and discount factor (γ) using a continuous four-rooms environment. Addition of temporal abstraction ($k > 1$) boosts performance, whereas increasing d or γ alone does not.

5.3 Temporal Abstraction and Effective Rank

Figure 2 shows how increasing the temporal abstraction step k reshapes the SR's singular value spectrum and reduces its effective rank. Across both discrete and continuous settings, larger k accelerates the decay of tail singular values, concentrating energy in dominant components relevant for long-horizon control. This aligns the SR with the low-rank inductive bias of FB, improving representation quality.

However, excessive abstraction is detrimental. As stable rank and spectral entropy approach their minima, task-relevant dynamics are lost. This reflects the theoretical trade-off between spectral compression and bias from action repetition. Empirically, Figure 4a shows performance degrading beyond an optimal k.

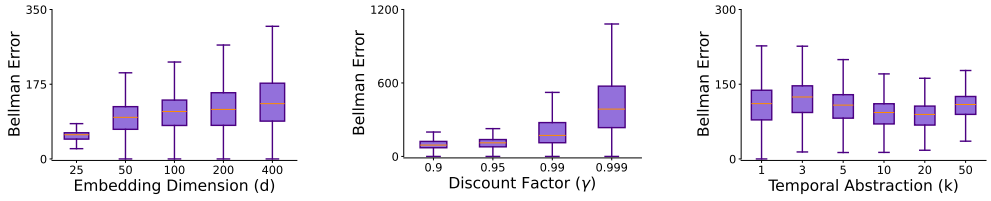
5.4 Temporal Abstraction and Embedding Dimension of FB

In principle, increasing the embedding dimension d should improve SR approximation (Blier et al., 2021; Touati & Ollivier, 2021). In continuous settings, however, this does not translate into better performance. Figure 5a shows that larger d increases Bellman error, while Figure 4b shows no performance gain without temporal abstraction ($k = 1$), even when scaling d from 25 to 400.

This behavior is consistent with our spectral analysis: higher capacity encourages fitting high-frequency components of a high-rank SR, which are hard to predict and amplify errors under bootstrapping. In contrast, temporal abstraction improves performance by reducing the effective rank of the target, simplifying the learning problem. Additional gains can be obtained by tuning d once k is fixed.

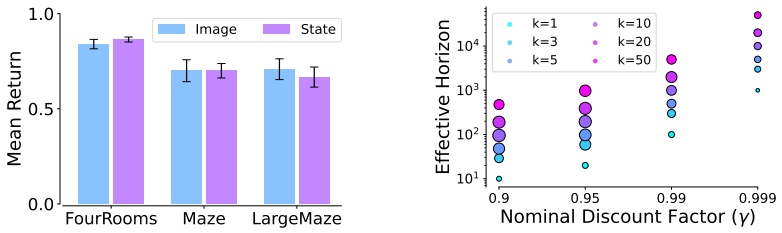
5.5 Spectral Dynamics: Discounting vs. Temporal Abstraction

A low-rank SR concentrates spectral energy in a small set of dominant components associated with long-horizon dynamics. As $\gamma \to 1$, these components are increasingly amplified, yielding stronger



(a) (b) (c)

Figure 5: **Ablations: Bellman error.** Increasing the embedding dimension (a) or discount factor (b) without using temporal abstraction ($k = 1$) leads to an increase in the Bellman error. Increasing k (c) does not increase the Bellman error.



(a) (b)

Figure 6: **Ablations: Input type and effective discount factor** (a): Image and State inputs use CNN and RBF encodings, respectively. (b): A higher return (larger radius) for a similar task horizon can be achieved by combining a lower nominal γ with a higher k. The effective horizon (y-axis) is computed as $\frac{1}{(1-\gamma^{1/k})}$, expressing the repeat-MDP horizon in environment-frame units.

spectral concentration but also reduced training stability. This effect is reflected in Figure 5b, where the absolute Bellman error grows with γ.

While normalizing the Bellman residual by the magnitude of the Q-values reverses this trend by compensating for the scaling of SR values, the optimization dynamics are governed by the absolute residual. Consequently, the increase in absolute Bellman error at large γ leads to higher gradient variance and a weaker effective contraction, which in turn degrades training stability. A more detailed comparison between relative and absolute Bellman errors is provided in Appendix D.4.

Discounting and temporal abstraction modify the spectrum in fundamentally different ways. Increasing γ amplifies existing components, including high-frequency ones, and degrades conditioning. In contrast, increasing k smooths the dynamics by attenuating sub-dominant, high-frequency components while preserving the steady-state structure.

Although both reduce effective rank, their behavior differs sharply. Discount-driven compression is abrupt and unstable, often causing collapse in stable rank and entropy. Temporal abstraction instead induces a controlled spectral decay: the effective rank drops quickly for small k and then stabilizes, maintaining higher spectral entropy. This enables a structured simplification of the predictive manifold without the instability of near-unity discounting.

6 A Recipe for Effective Forward-Backward Representations

Our results show that optimal performance arises from combining moderate discounting with temporal abstraction, rather than simply maximizing γ. Lower γ improves stability but shortens the effective horizon; this can be compensated by increasing the action-repeat factor k.

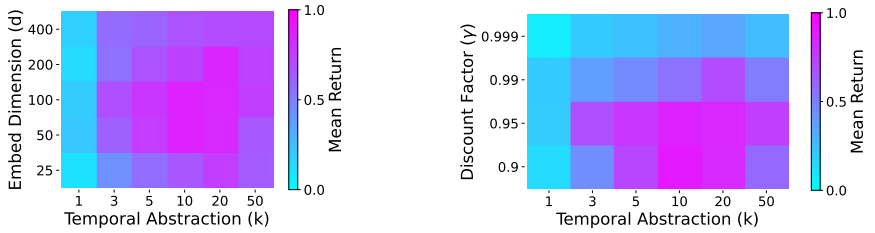
We distinguish between the nominal discount γ in the k-repeat MDP and the effective discount γ_{eff} in the original environment, related by $\gamma_{\text{eff}} = \gamma^{1/k}$. For a fixed task horizon, combining lower γ with larger k consistently outperforms standard high-discount settings (Figure 6b).

Figure 7 further shows that moderate temporal abstraction ($k \in [5, 10]$) acts as a robust regularizer across embedding dimensions and discount factors. Although the optimal k is environment-dependent, its inclusion yields consistent performance gains.

7 Conclusion

We identify a key mismatch between the low-rank inductive bias of FB and the inherently high-rank structure of the SR in continuous domains. The SR exhibits a heavy spectral tail of high-frequency components that are difficult to approximate and amplify errors under bootstrapping.

We address this mismatch by introducing temporal abstraction as a spectral regulator. Action repetition acts analogously to a low-pass filter, attenuating high-frequency components while preserving steady-state dynamics, thereby reducing the effective rank of the SR. This yields a simpler and more learnable



(a) (b)

Figure 7: **Ablation: Temporal Abstraction vs. Embedding Dimension vs. Discount Factor** Temporal abstraction offers a considerable boost in performance even with a moderate number of steps k. Performance is highest in the magenta region around $k \in [5, 10]$ across moderate γ and d. After the introduction of temporal abstraction, the performance is less sensitive to variations in the embedding dimension *(a)* than to changes in the discount factor *(b)*, especially as $\gamma \to 1$.

target for FB representations. Empirically, this spectral smoothing stabilizes learning and improves performance, even in high-discount regimes where standard FB struggles.

More broadly, our results suggest shifting focus from increasing model capacity to shaping the spectral structure of the underlying dynamics. Temporal abstraction serves as a practical tool for this purpose, enabling more stable and scalable predictive representations.

8 Limitations and Future Work

While our study uses continuous maze navigation to isolate the spectral effects of temporal abstraction, several research avenues remain. First, while these environments provide a controlled testbed for analyzing effective rank, generalizing our findings to domains with complex contact dynamics—such as locomotion or dexterous manipulation—is a primary direction for future work.

Second, we focus on action repetition as a fundamental form of temporal abstraction. More sophisticated frameworks, such as options or learned skills, may induce complex spectral transformations beyond the uniform attenuation studied here. Extending our analysis to adaptive abstractions could further clarify how hierarchical structures regularize representation learning.

Third, our results highlight an inherent trade-off between spectral stability and temporal resolution. As formalized in Definition 4.2, the smoothing that facilitates tractable learning also introduces a bias that limits resolution of high-frequency dynamics. This approach may therefore be less suitable for tasks requiring near-instantaneous reactive control.

Finally, while we consider online interaction in moderate dimensions, scaling to high-dimensional observations or offline settings (Sikchi et al., 2025; Tirinzoni et al., 2025) presents a compelling challenge. Investigating how temporal abstraction improves spectral conditioning in fixed datasets could significantly enhance the robustness of zero-shot generalization in offline reinforcement learning.

Acknowledgments

This work was funded by the Carl Zeiss Foundation through the ReScaLe project.

Broader Impact

This work advances the understanding of Forward–Backward representations, a general framework with potential applications across machine learning and robotics. We do not identify any immediate or specific societal risks beyond those broadly associated with these fields.

References

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4058–4068, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

André Biedenkapp, Raghu Rajan, Frank Hutter, and Marius Lindauer. Temporl: Learning when to act. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139, pp. 914–924, 2021.

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *CoRR*, abs/2101.07123, 2021. URL https://arxiv.org/abs/2101.07123.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Comput.*, 5(4):613–624, July 1993. ISSN 0899-7667. DOI: 10.1162/neco.1993.5.4.613. URL https://doi.org/10.1162/neco.1993.5.4.613.

Bastien Dubail, Stefan Stojanovic, and Alexandre Proutière. Shift before you learn: Enabling low-rank representations in reinforcement learning. *arXiv preprint arXiv:2509.05193*, 2025.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. URL https://api.semanticscholar.org/CorpusID:10163399.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge; New York, 2nd edition, 2013. ISBN 9780521839402.

Tejas D. Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J. Gershman. Deep successor reinforcement learning. *ArXiv*, abs/1606.02396, 2016. URL https://api.semanticscholar.org/CorpusID:11965834.

Aravind S. Lakshminarayanan, Sahil Sharma, and Balaraman Ravindran. Dynamic action repetition for deep reinforcement learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 2133–2139. AAAI Press, 2017.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2295–2304. JMLR.org, 2017a.

Marlos C. Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. *ArXiv*, abs/1710.11089, 2017b. URL https://api.semanticscholar.org/CorpusID:3300406.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: evaluation protocols and open problems for general agents (extended abstract). In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 5573–5577. AAAI Press, 2018. ISBN 9780999241127.

Sridhar Mahadevan. Proto-value functions: developmental reinforcement learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 553–560, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. DOI: 10.1145/1102351.1102421. URL https://doi.org/10.1145/1102351.1102421.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025.

Dikshant Shehmar, Matthew Schlegel, Matthew E. Taylor, and Marlos C. Machado. Laplacian representations for decision-time planning. *CoRR*, abs/2602.05031, 2026.

Harshit S. Sikchi, Andrea Tirinzoni, Ahmed Touati, Yingchen Xu, Anssi Kanervisto, Scott Niekum, Amy Zhang, Alessandro Lazaric, and Matteo Pirotta. Fast adaptation with behavioral foundation models. *ArXiv*, abs/2504.07896, 2025. URL https://api.semanticscholar.org/CorpusID:277667156.

Gilbert W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic, Boston, 1990. ISBN 0126702306 9780126702309. URL https://www.worldcat.org/title/matrix-perturbation-theory/oclc/908946968.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

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):181–211, 1999. ISSN 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(99)00052-1. URL https://www.sciencedirect.com/science/article/pii/S0004370299000521.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control via behavioral foundation models. *ArXiv*, abs/2504.11054, 2025. URL https://api.semanticscholar.org/CorpusID:277787058.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NeurIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=MYEap_OcQI.

Jingwei Zhang, Jost Tobias Springenberg, Joschka Boedecker, and Wolfram Burgard. Deep reinforcement learning with successor features for navigation across similar environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2371–2378. IEEE Press, 2017. DOI: 10.1109/IROS.2017.8206049. URL https://doi.org/10.1109/IROS.2017.8206049.

A Hyperparameters and Implementation Details

We summarize the key hyperparameters used for training the forward-backward (FB) representation in Table 1. These settings were kept fixed across experiments unless stated otherwise. The k-step action repetition is implemented via a wrapper over the environment. In other words, the agent will only interact with the k-repeat MDP and will not have access to the intermediate observations among the k steps. We use episodic return as a measure of performance for the agents. All environments provide a sparse (zero or one) reward. To get the final performance or on each validation step, each model is evaluated on 50 episodes with random start and goal resets. The latent vector z describing the task or goal is sampled during training using a 50-50 mix of sampling from a normal distribution or from a random visited state in the replay buffer projected to the latent space using the backward network, similarly to Tirinzoni et al. (2025).

Table 1: Relevant hyperparameters used for training FB representation.

Training steps (gradient update)	1e6
Reward type	Sparse
Hidden Layers (Forward, Backward, Actor)	[256, 256]
Ensemble (Forward)	2
Learning Rate (Backward, Actor)	1e-6
Learning Rate (Forward)	1e-5
Batch size (state)	512
Batch size (image)	128
Replay Buffer size	1e6
FB Orthogonal Loss Coef.	1.0
z-latent: buffer data vs. random sampling ratio	0.5
z-latent: hold steps before resampling	10

B Proofs

In the following, we provide the proofs of this work. Note, that in Touati & Ollivier (2021), the reward embedding $z_R \coloneqq B^\top r\nu$ is weighted by a data distribution ν. For this work, we assume a uniform distribution and implicitly absorb its normalization constant into the scaling of B, simplifying the embedding to the matrix-vector product $z_R = B^\top r$.

B.1 Optimality Gap for FB Representations

The following theorem is a simplified refinement (Touati & Ollivier, 2021, Theorem 8) for our spectral analysis setting.

Theorem 3.1 (Optimality Gap for FB Representations)**.** *Let $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be a reward function such that $z_R = B^\top r$. The approximation error of the optimal Q-function is bounded by:*

$$\left\|F(\,\cdot\,,\,\cdot\,,\,z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2C_{\text{norm}}\,\|r\|_\infty}{(1-\gamma)}\|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2. \tag{2}$$

Proof. Applying (Touati & Ollivier, 2021, Theorem 8) to our setting we have

$$\left\|F(\,\cdot\,,\,\cdot\,,\,z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2\|r\|_A}{(1-\gamma)}\sup_{s,a}\|\hat{M}^{z_R}(s,a,\,\cdot\,,\,\cdot\,) - M^{\pi_{z_R}}(s,a,\,\cdot\,,\,\cdot\,)\|_B,$$

where the norms $\|\,\cdot\,\|_A$ on functions and $\|\,\cdot\,\|_B$ on (signed) measures must satisfy the duality compatibility $|\langle f,\mu\rangle| \leq \|f\|_A\|\mu\|_B$ for all f,μ. Choosing $\|\cdot\|_A = \|\cdot\|_\infty$ and $\|\cdot\|_B = \|\cdot\|_2$, Hölder's inequality combined with $\|\mu\|_1 \leq \sqrt{|\mathcal{S}||\mathcal{A}|}\,\|\mu\|_2$ yields $|\langle f,\mu\rangle| \leq \|f\|_\infty\|\mu\|_1 \leq C_{\text{norm}}\,\|f\|_\infty\,\|\mu\|_2$ with $C_{\text{norm}} = \sqrt{|\mathcal{S}||\mathcal{A}|}$. Given this, we have

$$\begin{aligned}\left\|F(\,\cdot\,,\,\cdot\,,\,z_R)^\top z_R - Q^\star\right\|_\infty &\leq \frac{2\,C_{\text{norm}}\,\|r\|_\infty}{(1-\gamma)}\sup_{s,a}\|\hat{M}^{z_R}(s,a,\,\cdot\,,\,\cdot\,) - M^{\pi_{z_R}}(s,a,\,\cdot\,,\,\cdot\,)\|_2 \\ &\leq \frac{2\,C_{\text{norm}}\,\|r\|_\infty}{(1-\gamma)}\|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2.\end{aligned}$$

This ends the proof. □

B.2 Spectral Bound for Optimality Gap of k-repeat FB Representations

We first derive the matrix form of the action-repeat policy-induced transition matrix used by both Lemma 4.1 and Lemma 4.2.

For a fixed action a_q, let $P_{a_q} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the state-transition dynamics under that action, and let $\pi_{s_p} \in \mathbb{R}^{1\times|\mathcal{A}|}$ represent the row vector of policy probabilities for a given state s_p, i.e.,

$$P_{a_q} = \begin{bmatrix} P(s_1 \mid s_1, a_q) & \dots & P(s_{|\mathcal{S}|} \mid s_1, a_q) \\ \vdots & \ddots & \vdots \\ P(s_1 \mid s_{|\mathcal{S}|}, a_q) & \dots & P(s_{|\mathcal{S}|} \mid s_{|\mathcal{S}|}, a_q) \end{bmatrix}, \quad \pi_{s_p} = \begin{bmatrix}\pi(a_1 \mid s_p) & \dots & \pi(a_{|\mathcal{A}|} \mid s_p)\end{bmatrix}.$$

Stacking the per-action matrices into a block-diagonal matrix $\mathbf{P}_\mathcal{A}$, we can express the transition matrix $\widetilde{P}^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ as a product of action-repetition and policy-mapping components:

$$\widetilde{P}^\pi = K\,\mathbf{P}_\mathcal{A}^k\,E\,\tilde{\pi}, \tag{A1}$$

where

$$\mathbf{P}_\mathcal{A}^k = \begin{bmatrix} P_{a_1}^k & & 0 \\ & \ddots & \\ 0 & & P_{a_{|\mathcal{A}|}}^k \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|} \quad \text{and} \quad \tilde{\pi} = \begin{bmatrix} \pi_{s_1} & & 0 \\ & \ddots & \\ 0 & & \pi_{s_{|\mathcal{S}|}} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}\times\mathcal{A}|}.$$

The matrix $K \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is a commutation matrix that reorders the state-action product space from a state-major to an action-major indexing scheme, and $E \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ is a broadcasting

matrix that lifts a vector from $\mathbb{R}^{|\mathcal{S}|}$ to $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ by replicating each state coordinate $|\mathcal{A}|$ times. In the decomposition (A1), $\mathbf{P}_{\mathcal{A}}^{k}$ captures the k-step transitions under action repetition, while $\widetilde{\pi}$ maps the policy within the state-action space. Intuitively, the system first evolves for k steps under the same action, after which the next action is selected according to the policy without execution.

Before stating the lemma we introduce a diagonalizability assumption on the joint chain.

Assumption B.1 (Diagonalizability of joint dynamics)**.** *The policy-induced transition matrix $\widetilde{P}^{\pi}$ of the action-repeat MDP is diagonalizable over $\mathbb{C}$, i.e. $\widetilde{P}^{\pi} = S\Lambda S^{-1}$ for some invertible matrix S.*

This is a standard assumption in spectral analyses of Markov chains and matrix perturbation theory (Horn & Johnson, 2013; Stewart & Sun, 1990) and is generic: matrices with distinct eigenvalues are dense in $\mathbb{R}^{n\times n}$ (Horn & Johnson, 2013, Theorem 2.4.7.1). We therefore expect it to hold in essentially most discrete environments of practical interest, since diagonalizability fails only when two or more eigenvalues coincide and their eigenvectors fail to span the corresponding joint eigenspace, an exact algebraic degeneracy broken by any stochasticity or asymmetry in the transition dynamics. The tightness of the bound is controlled by $\kappa(S) = \|S\|_2\|S^{-1}\|_2$, which equals 1 when $\widetilde{P}^{\pi}$ is normal and grows as $\widetilde{P}^{\pi}$ approaches a defective matrix.

Lemma 4.1 (Spectral Bound of Optimality Gap for k-repeat FB Representations)**.** *Given a reward function $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, let $\widetilde{r}(s,a) \coloneqq \mathbb{E}_P[\sum_{t=0}^{k-1}\gamma^t r(s_t,a) \mid s_0 = s]$ be the corresponding k-step expected reward function, and define $z_{\widetilde{r}} \coloneqq B^{\top}\widetilde{r}$. Under Assumption B.1, let (F,B) be an FB representation with dimension d. Then the error in approximating the original optimal action-value function $Q^{\star}$ is bounded by:*

$$\left\|F(\cdot\,,\cdot\,,z_{\widetilde{r}})^{\top}z_{\widetilde{r}} - Q^{\star}\right\|_{\infty} \le \epsilon_{\mathrm{repeat}}(k) + \frac{2C_{\mathrm{norm}}\,\|r\|_{\infty}}{1-\gamma}\left(\widetilde{\epsilon}_{\mathrm{real}}(\widetilde{r}) + \frac{C_{\mathrm{SF}}}{1-\gamma^{k}|\lambda_{d+1}(\widetilde{P}^{\pi})|}\right).$$

Proof. The proof bounds the spectral truncation error of the action-repeat successor representation $\widetilde{M}^{\pi}$ in terms of $|\lambda_{d+1}(\widetilde{P}^{\pi})|$ (Step 1), the k-step reward magnitude (Step 2), and combines these into the final bound (Step 3).

Step 1: Spectral bound of the discounted infinite horizon. The successor representation $\widetilde{M}^{\pi}$ for the action-repeat MDP is defined by the discounted sum of future transitions $\widetilde{M}^{\pi} \coloneqq \sum_{t=0}^{\infty}(\gamma^{k}\widetilde{P}^{\pi})^{t}$. Since $\widetilde{P}^{\pi}$ is a row-stochastic matrix, its spectral radius satisfies $\rho(\widetilde{P}^{\pi}) = 1$ and thus with discounting we have $\rho(\gamma^{k}\widetilde{P}^{\pi}) = \gamma^{k} < 1$. Thus, the Neumann series converges to the matrix inverse:

$$\widetilde{M}^{\pi} = (I - \gamma^{k}\widetilde{P}^{\pi})^{-1}.$$

Since $\widetilde{P}^{\pi} = S\Lambda S^{-1}$ (Assumption B.1), the successor representation of the action-repeat MDP (Definition 4.1, with discount γ^{k}) is

$$\widetilde{M}^{\pi} = (I - \gamma^{k}\widetilde{P}^{\pi})^{-1} = S\,(I - \gamma^{k}\Lambda)^{-1}\,S^{-1} = S\,\mathrm{diag}\!\left(\frac{1}{1-\gamma^{k}\lambda_i}\right)S^{-1}.$$

Let $\Phi = (I - \gamma^{k}\Lambda)^{-1} = \mathrm{diag}(1/(1-\gamma^{k}\lambda_i))$. By submultiplicativity of singular values (Horn & Johnson, 2013):

$$\sigma_{d+1}(\widetilde{M}^{\pi}) = \sigma_{d+1}(S\Phi S^{-1}) \le \|S\|_2\cdot\sigma_{d+1}(\Phi)\cdot\|S^{-1}\|_2 = \kappa(S)\cdot\sigma_{d+1}(\Phi).$$

Let $\lambda_i = \lambda_i(\widetilde{P}^{\pi})$ be the i-th eigenvalue of $\widetilde{P}^{\pi}$, ordered by nonincreasing modulus. Since Φ is diagonal, its singular values are the moduli of its diagonal entries, which we denote $\phi_i \coloneqq 1/|1-\gamma^{k}\lambda_i|$ for $i = 1,\dots,n$. Note that, as λ_i may be complex the sequence of ϕ_i is not sorted by modulus and we have in general $\sigma_i(\Phi) \neq \phi_i$ but $\sigma_j(\Phi) = [(j\text{-th largest}\{\phi_i\}]$. Since $\widetilde{P}^{\pi}$ is row-stochastic we have $|\lambda_i| \le 1$, so $\gamma^{k}|\lambda_i| < 1$. Applying the reverse triangle inequality to each denominator gives the pointwise bound

$$\phi_i = \frac{1}{|1-\gamma^{k}\lambda_i|} \;\le\; \frac{1}{1-\gamma^{k}|\lambda_i|} \;=:\; \tilde{\phi}_i,$$

where the dominating sequence $\tilde{\phi}_i$ is monotone increasing in $|\lambda_i|$. Since $|\lambda_i|$ is sorted nonincreasingly, the $(d{+}1)$-th largest element of $\{\tilde{\phi}_i\}_{i=1}^n$ is exactly $\tilde{\phi}_{d+1} = 1/(1-\gamma^k|\lambda_{d+1}|)$.

Combining the pointwise bound $\phi_i \le \widetilde{\phi}_i$ with the monotonicity of $\widetilde{\phi}_i$ in $|\lambda_i|$, we obtain

$$\sigma_{d+1}(\Phi) \;=\; \big[(d{+}1)\text{-th largest of } \{\phi_i\}\big] \;\le\; \big[(d{+}1)\text{-th largest of } \{\widetilde{\phi}_i\}\big] \;=\; \widetilde{\phi}_{d+1} \;=\; \frac{1}{1-\gamma^k|\lambda_{d+1}|}.$$

Combining with the submultiplicativity bound, the SVD truncation error can be bounded by

$$\sigma_{d+1}(\widetilde{M}^\pi) \le \kappa(S)\cdot\sigma_{d+1}(\Phi) \le \frac{\kappa(S)}{1-\gamma^k|\lambda_{d+1}(\widetilde{P}^\pi)|}.$$

Step 2: Bound reward $\widetilde{r}$. By definition, the expected k-step reward is given by $\widetilde{r}(s,a) = \mathbb{E}_P\left[\sum_{t=0}^{k-1}\gamma^t r(s_t,a) \mid s_0 = s\right]$. Bounding the reward at each step with $\|r\|_\infty$, we get:

$$\|\widetilde{r}\|_\infty \le \sum_{t=0}^{k-1}\gamma^t\|r\|_\infty = \frac{1-\gamma^k}{1-\gamma}\|r\|_\infty.$$

Step 3: Complete the proof. Using the triangle inequality and Definition 4.2, we first separate the action-repeat approximation error:

$$\begin{aligned}\left\|F(\cdot,\cdot,z_{\widetilde{r}})^\top z_{\widetilde{r}} - Q^\star\right\|_\infty &\le \left\|F(\cdot,\cdot,z_{\widetilde{r}})^\top z_{\widetilde{r}} - \widetilde{Q}^\star\right\|_\infty + \|\widetilde{Q}^\star - Q^\star\|_\infty \\ &= \left\|F(\cdot,\cdot,z_{\widetilde{r}})^\top z_{\widetilde{r}} - \widetilde{Q}^\star\right\|_\infty + \epsilon_{\text{repeat}}(k).\end{aligned}$$

Next, applying Theorem 3.1 to the first term and substituting our reward bound from Step 2:

$$\begin{aligned}&\le \epsilon_{\text{repeat}}(k) + \frac{2\,C_{\text{norm}}\,\|\widetilde{r}\|_\infty}{1-\gamma^k}\|\hat{M}^{z_{\widetilde{r}}} - \widetilde{M}^{\pi_{z_{\widetilde{r}}}}\|_2 \\ &\le \epsilon_{\text{repeat}}(k) + \frac{2\,C_{\text{norm}}\,\|r\|_\infty}{1-\gamma}\|\hat{M}^{z_{\widetilde{r}}} - \widetilde{M}^{\pi_{z_{\widetilde{r}}}}\|_2.\end{aligned}$$

Finally, bounding the model error by the realizability error (Definition 3.1) and the spectral truncation properties established in Step 1:

$$\begin{aligned}&\le \epsilon_{\text{repeat}}(k) + \frac{2\,C_{\text{norm}}\,\|r\|_\infty}{1-\gamma}\left(\widetilde{\epsilon}_{\text{real}}(\widetilde{r}) + \sigma_{d+1}(\widetilde{M}^{\pi_{z_{\widetilde{r}}}})\right) \\ &\le \epsilon_{\text{repeat}}(k) + \frac{2\,C_{\text{norm}}\,\|r\|_\infty}{1-\gamma}\left(\widetilde{\epsilon}_{\text{real}}(\widetilde{r}) + \frac{\kappa(S)}{1-\gamma^k|\lambda_{d+1}(\widetilde{P}^\pi)|}\right).\end{aligned}$$

Defining $C_{\text{SF}} \coloneqq \kappa(S)$ completes the proof. □

B.3 Eigenvalue Contraction under Action Repetition

Lemma 4.1 bounds the truncation error in terms of $|\lambda_{d+1}(\widetilde{P}^\pi)|$, but does not say how this eigenvalue depends on the action-repeat horizon k. We now show that, under a structural condition on the per-action transition matrices, $|\lambda_{d+1}(\widetilde{P}^\pi)|$ contracts when increasing k, exposing the explicit role of action repetition. The argument works from the block-diagonal factorization (A1) and requires diagonalizability of each block.

Assumption B.2 (Diagonalizability of action blocks)**.** *For all actions $a \in \mathcal{A}$ the corresponding state transition matrix P_a is diagonalizable over $\mathbb{C}$, i.e. $P_a = U_a\Lambda_a U_a^{-1}$.*

As with Assumption B.1, this is generic: matrices with distinct eigenvalues are dense in $M_n(\mathbb{R})$ (Horn & Johnson, 2013, Thm. 2.4.7.1). Note that diagonalizing the per-action blocks P_a is a stronger requirement than diagonalizing the joint matrix $\widetilde{P}^\pi$, since the proof relies on the block-diagonal eigendecomposition of $\mathbf{P}_\mathcal{A}$ being matched with $\widetilde{\pi}$.

Lemma 4.2 (Eigenvalue Contraction under Action Repetition)**.** *Under Assumption B.2, the* $(d{+}1)$*-th largest absolute eigenvalue of* $\widetilde{P}^{\pi}$ *contracts exponentially in* k*:*

$$|\lambda_{d+1}(\widetilde{P}^{\pi})| \;\le\; C_{\mathrm{rep}}\,|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^{k}.$$

Proof. We work from the factorization (A1). Per Assumption B.2, each transition matrix P_a is diagonalizable as $P_a = U_a \Lambda_a U_a^{-1}$, so $\mathbf{P}_{\mathcal{A}} = U \Lambda_{\mathcal{A}} U^{-1}$ with the block-diagonal eigenvector and eigenvalue matrices

$$U = \mathrm{diag}(U_{a_1}, \ldots, U_{a_{|\mathcal{A}|}}), \quad \Lambda_{\mathcal{A}} = \mathrm{diag}(\Lambda_{a_1}, \ldots, \Lambda_{a_{|\mathcal{A}|}}).$$

Substituting into (A1) and raising to the k-th power gives

$$\widetilde{P}^{\pi} = KU\Lambda_{\mathcal{A}}^{k}U^{-1}E\tilde{\pi}.$$

To extract the $(d{+}1)$-th eigenvalue, we partition the 1-step spectrum $\Lambda_{\mathcal{A}}$ into the d largest eigenvalues (Λ_{slow}) and the remaining fast-mixing eigenvalues (Λ_{fast}):

$$\Lambda_{\mathcal{A}} = \begin{pmatrix} \Lambda_{\mathrm{slow}} & 0 \\ 0 & \Lambda_{\mathrm{fast}} \end{pmatrix}.$$

By definition, the largest absolute value in Λ_{fast} is exactly $|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|$. We now decompose the full system into a rank-d matrix (M_d) and a fast-mixing error matrix (E_k):

$$\widetilde{P}^{\pi} = \underbrace{KU \begin{pmatrix} \Lambda_{\mathrm{slow}}^{k} & 0 \\ 0 & 0 \end{pmatrix} U^{-1}E\tilde{\pi}}_{=:M_d} + \underbrace{KU \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_{\mathrm{fast}}^{k} \end{pmatrix} U^{-1}E\tilde{\pi}}_{=:E_k}.$$

Let S be the eigenvector matrix of $\widetilde{P}^{\pi}$ (Assumption B.1 is invoked here only to make $\kappa(S)$ well-defined; the contraction itself relies only on Assumption B.2). To bound the $(d+1)$-th eigenvalue of the system, we rely on global eigenvalue matching bounds for diagonalizable matrices. Let $N = |\mathcal{S}||\mathcal{A}|$ be the dimension of the space. Theorem 3.3 in Stewart & Sun (1990) establishes the existence of an optimal permutation $\tau^{\star}$ matching the spectra of $\widetilde{P}^{\pi}$ and M_d that minimizes the maximum deviation between paired eigenvalues:

$$\min_{\tau} \max_{i} |\lambda_i(\widetilde{P}^{\pi}) - \lambda_{\tau(i)}(M_d)| \le (2N-1)\kappa(S)\|E_k\|_2.$$

Because M_d has rank at most d, it possesses at least $N-d$ zero eigenvalues, so M_d has at most d nonzero eigenvalues. By the pigeonhole principle applied to the top $d{+}1$ eigenvalues of $\widetilde{P}^{\pi}$ (those with modulus nonsmaller than $|\lambda_{d+1}(\widetilde{P}^{\pi})|$), at least one index $i^{\star} \in \{1, \ldots, d{+}1\}$ must satisfy $\lambda_{\tau^{\star}(i^{\star})}(M_d) = 0$. Combining the matching bound at $i^{\star}$ with the modulus ordering:

$$|\lambda_{d+1}(\widetilde{P}^{\pi})| \;\le\; |\lambda_{i^{\star}}(\widetilde{P}^{\pi})| \;=\; |\lambda_{i^{\star}}(\widetilde{P}^{\pi}) - \lambda_{\tau^{\star}(i^{\star})}(M_d)| \;\le\; (2N-1)\kappa(S)\|E_k\|_2,$$

where the first inequality holds because $i^{\star} \le d{+}1$ and the eigenvalues are ordered by nonincreasing modulus.

Using the submultiplicativity of the spectral norm, we bound $\|E_k\|_2$:

$$\begin{aligned} \|E_k\|_2 &\le \|K\|_2 \cdot \|U\|_2 \cdot \|\Lambda_{\mathrm{fast}}^{k}\|_2 \cdot \|U^{-1}\|_2 \cdot \|E\tilde{\pi}\|_2 \\ &\le 1 \cdot \kappa(U) \cdot |\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^{k} \cdot \sqrt{|\mathcal{A}|}, \end{aligned}$$

where we use the fact that K is a permutation matrix ($\|K\|_2 = 1$), $\kappa(U) = \|U\|_2\|U^{-1}\|_2$, and we bound $\|E\tilde{\pi}\|_2 \le \|E\|_2 \le \sqrt{|\mathcal{A}|}$. Substituting this back yields the explicit bound on the $(d+1)$-th eigenvalue of the system:

$$|\lambda_{d+1}(\widetilde{P}^{\pi})| \le \underbrace{(2|\mathcal{S}||\mathcal{A}|-1)\kappa(S)\kappa(U)\sqrt{|\mathcal{A}|}}_{=:C_{\mathrm{rep}}} \cdot |\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^{k},$$

which is the claim of the lemma. □

A structural artifact of the block-diagonal factorization (A1) is that each block P_a is row-stochastic, so $\mathbf{P}_{\mathcal{A}}$ has at least $|\mathcal{A}|$ unit eigenvalues. Consequently, $|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})| < 1$ requires $d \geq |\mathcal{A}|$; otherwise Lemma 4.2 is vacuous because $|\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^k = 1$ for all k. This constraint is specific to the block diagonal proof strategy used here.

Combining this with Lemma 4.1 via direct substitution yields the explicit form of the spectral error term as

$$\frac{C_{\mathrm{SF}}}{1 - \gamma^k C_{\mathrm{rep}} |\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^k},$$

which is meaningful whenever $C_{\mathrm{rep}} |\lambda_{d+1}(\mathbf{P}_{\mathcal{A}})|^k < \gamma^{-k}$. Under stronger structural assumptions such as orthogonality of the action block matrices P_a the same proof strategy would potentially tighten C_{rep} to $\sqrt{|\mathcal{A}|}$.

C Spectral Metrics

We define two spectral metrics to quantify the effective rank of the SR.

Stable Rank. The stable rank captures the concentration of spectral energy relative to the dominant singular direction. For a matrix M, it is defined as:

$$\mathrm{SRank}(M) = \frac{\|M\|_F^2}{\|M\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2},$$

where $\{\sigma_i\}$ are the singular values of M. Lower values indicate stronger concentration in leading components and thus greater low-rank structure.

Normalized Spectral Entropy. Normalized spectral entropy (NSE) measures how evenly spectral energy is distributed:

$$\mathrm{NSE}(M) = \frac{-\sum_i p_i \log p_i}{\log(\beta)}, \quad p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2},$$

where β denotes the number of singular values. This normalization ensures $\mathrm{NSE}(M) \in [0, 1]$. Higher values correspond to a more diffuse spectrum, while lower values indicate concentration in a few modes.

Discrete Setting. In discrete environments, we compute both metrics directly on the exact SR matrix M^π by performing singular value decomposition (SVD) to obtain $\{\sigma_i\}$.

Continuous Setting. In continuous domains, where the exact SR is unavailable, we evaluate the metrics on an empirical approximation $\hat{M}^\pi = FB^\top$, constructed from transitions collected under the same exploration protocol used during training. In contrast to training—where a batch of latent embeddings is sampled—we use a single randomly sampled latent embedding shared across all transitions, yielding an estimate of $\hat{M}^\pi$ for a fixed (random) goal.

To mitigate scale drift arising from variations in embedding norms, we apply a row-wise softmax normalization such that each row sums to $\frac{1}{1-\gamma}$. We then compute the singular values of the normalized matrix via SVD and evaluate the spectral metrics as in the discrete case. All reported results correspond to averages of the metrics over all trained models using random seeds (see Figure 2).

D Extra Plots and Ablations

D.1 Overall effects of k, γ, and d

Figure 8 provides an overview of the effect of the three main hyperparameters of FB on final episodic return of the Four-Rooms continuous environment. Two values are of particular importance, action-repetition (k=1) and nominal discount factor (γ=0.999). In both cases, the performance suffers significantly regardless of the values of other hyperparameters. In the case of k=1 or no temporal abstraction, FB networks find it challenging to learn a good representation due to the presence of unpredictable high-frequency dynamical modes. In the case high discount factor, γ=0.999, a good representation cannot be achieved as the representation rank approaches singularity. The learning is less sensitive overall to the value of the embedding dimension d.

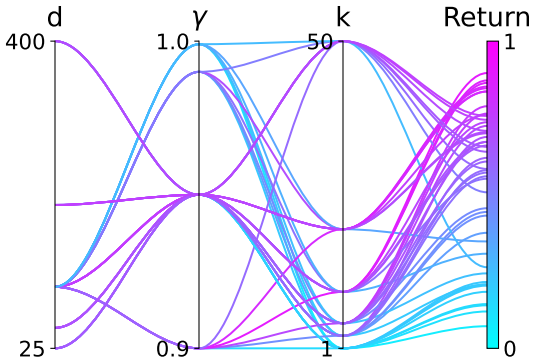
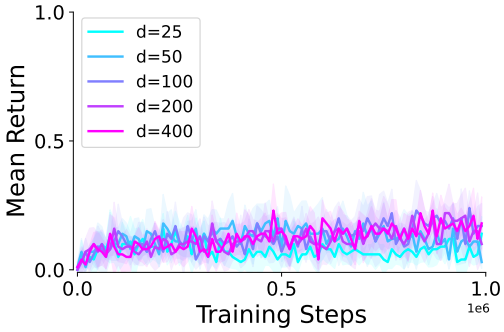


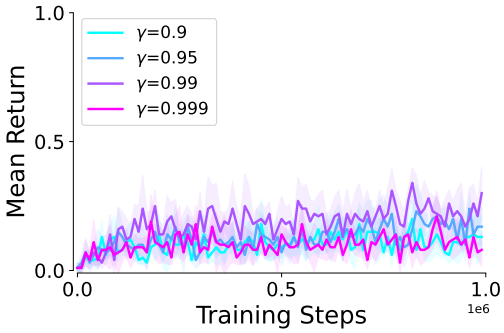
Figure 8: **Relationship between the main hyperparameters and episodic return.** This plot gives an overview of the effect of different combinations of embedding dimension (d), discount factor (γ), and temporal abstraction k over all experiments. Noticeably, $k = 1$ or $\gamma = 0.999$ leads to poor performance in most combinations.

D.2 Training plots: Ablation of k, γ, and d

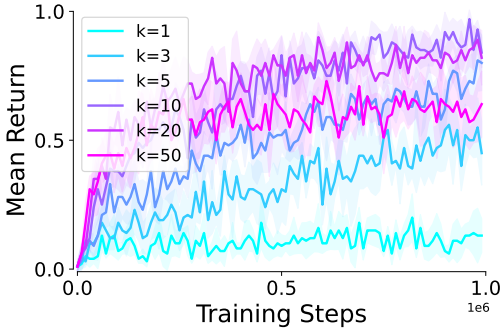
Figure 9, shows the performance of different combination of the main hyperparameters during training with the focus on the effect of introducing temporal abstraction. Figures 9a and 9b highlight that without temporal abstraction (k=1) varying the embedding dimension d or the discount factor γ yields no significant improvement. Figure 9c, on the other hand shows that even a small level of temporal abstraction (k=3) can lead to a significant boost in performance. The figure also shows the limitation of the temporal abstraction where a large temporal abstraction (k=50) can start to have negative impact on the performance, by oversimplification of the SR representation and removing dynamical modes that are useful for the navigation task.



(a) γ=0.95, k=1



(b) d=100, k=1



(c) d=100, γ=0.95

Figure 9: **Ablation: Training plots.** Increasing embedding dimension d or discount factor γ without increasing the temporal abstraction does not yield a meaningful increase in performance.

D.3 SR and its Q-function for discrete and continuous Four-Rooms environment

Figure 10 shows a more complete picture of SR and its associated Q function (mean over cardinal action directions). The *Baseline* shows the SR and Q using no temporal abstraction (k=1), and moderate discount factor (γ=0.95). For the continuous settings SR is calculated via FB with embedding dimension (d=100).

In discrete settings (top two rows), a low-rank structure can be achieved in three ways: 1) SVD with a small rank *(rank = 4)*, 2) high discount factor *(γ=0.999)*, or 3) using temporal abstraction via action repetition *(k=10)*. In the absence of function approximation and bootstrapping all three paths lead to an overall similar result where a low-rank structure can remove the high frequency dynamical modes and create shared future topology (rooms, corridors,...) where states with similar reachability are grouped together and have similar values.

In continuous settings (bottom two rows), where SR and its associated Q are learned via FB using function approximation and bootstrapping, the results differ. To enforce a low-rank structure via the FB algorithm we reduce the embedding dimension from 100 to 25. The figure shows a small smoothing (grouping of states), but this is not nearly close to the effect of enforcing low-rank structure using SVD in the discrete setting. Increasing the discount factor (γ=0.999) and introducing temporal abstraction via action repetition (k=10) show more promise as they both help spread the SR and Q values to the neighboring rooms. However, a closer look at the Q values shows that only temporal abstraction can smoothly distribute the Q values as the states move away from the goal (start marker). The policy based on increased γ will be stuck in local maxima while the policy based on increased k can follow the Q gradients to the goal.

Figure 10: **Successor Representation (SR) and its Q-Function - Discrete and Continuous** This is a more complete picture of Figure 1. The Q-functions are derived from the same SR that is presented here. The *star marker* marks the starting state for SR and the goal state for the Q-function.

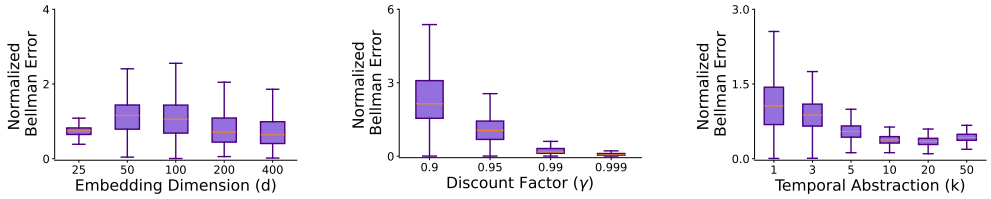
D.4 Absolute vs Relative Bellman Error

Figure 11 presents the normalized Bellman error corresponding to Figure 5, where the residuals are scaled by the magnitude of the Q-values. We observe that increasing the embedding dimension has a negligible effect on the relative Bellman error, whereas increasing the degree of temporal abstraction consistently reduces it.

This reduction becomes more pronounced as the discount factor γ increases, in contrast to the trend observed for the absolute Bellman error. Overall, the results reveal a clear divergence between these metrics at large γ: the relative Bellman error decreases, while the episodic return simultaneously deteriorates (Figure 4c).

We hypothesize that this discrepancy is driven by the growth of the absolute Bellman error. As $\gamma \to 1$, the scale of the successor representation increases proportionally to the effective horizon, $(1-\gamma)^{-1}$, which artificially attenuates the normalized error. However, optimization is governed by the absolute Bellman residual. Thus, larger absolute errors at high γ lead to increased gradient variance and a weaker contraction effect, resulting in training instability.

These findings suggest that the absolute Bellman error is a more reliable indicator of policy degradation than its normalized counterpart, as it more faithfully captures the intrinsic difficulty of function approximation in long-horizon regimes.

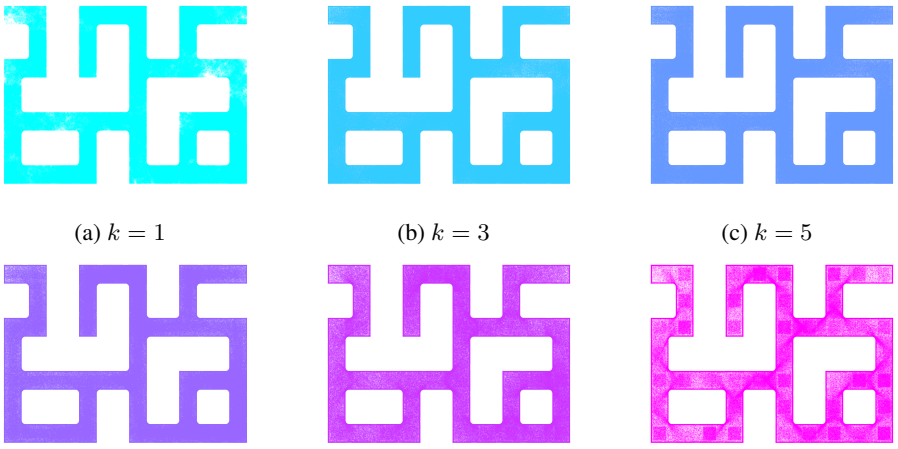


(a) (b) (c)

Figure 11: **Ablations: Normalized Bellman error.** Bellman errors are normalized by the Q values. The relative Bellman error decreases sharply as the discount factor is increased (b). However, this decrease in relative Bellman error does not translate to better performance (episodic return) as discussed in Section 5.

E Exploration Coverage

In order to verify that increasing action repetition did not significantly influence the exploration coverage of the state space, we plot the states visited during the training of agents with varying action repetition values for the LargeMaze environment. We use one interaction in the k-repeat environment per training step, hence each agent visits one million states during its training.



(a) $k = 1$ (b) $k = 3$ (c) $k = 5$

(d) $k = 10$ (e) $k = 20$ (f) $k = 50$

Figure 12: **Scatter plot of visited states during training.** Increasing the action repetition does not significantly affect the coverage of the space in the LargeMaze environment. Similar coverage pattern holds Four-Rooms and Maze environments. (omitted for brevity)

F Compute Resources

Each experiment (1M training steps) was conducted using a single GPU (NVIDIA GeForce RTX 2080 Ti), taking an average of 12 hours per experiment when training with state observations and an average of 18 hours when using image observations. Each experiment used 1.5GB of RAM when trained using state observations and 13GB of RAM when trained with image observations of size $(64 \times 64 \times 3)$ pixels.