LATENT-PREDICTIVE UNIVERSAL HORIZON MODELS FOR LONG HORIZON ZERO-SHOT RL

Anonymous authors
Paper under double-blind review

ABSTRACT

While behavior foundation models (BFMs) have achieved impressive zero-shot performance on various human locomotion tasks, these methods struggle on tasks that require longer horizon reasoning. Recent work has explored using geometric horizon models (GHMs) for planning on top of pre-trained BFMs to achieve longer horizon tasks; however, such methods require significant test-time compute, operate in raw observation space, and require a complex two-stage training process. In this work, we instead explore training horizon models in latent space and as an auxiliary loss *during* BFM training. In particular, we propose a novel latent-predictive Universal Horizon Model, which we learn jointly with a latent-predictive BFM with both models shaping the state- and task-encoding spaces. We demonstrate that this auxiliary representation learning objective is theoretically compatible with the primary TD-JEPA objective, with respect to the learned state and task encodings, and we demonstrate empirically that it can be beneficial for downstream zero-shot performance. Finally, we show that this UHM model can be used to perform model-based updates in the zero-shot RL setting, allowing us to stably train a latent-predictive BFM at discounts as high as $\gamma = 0.999$.

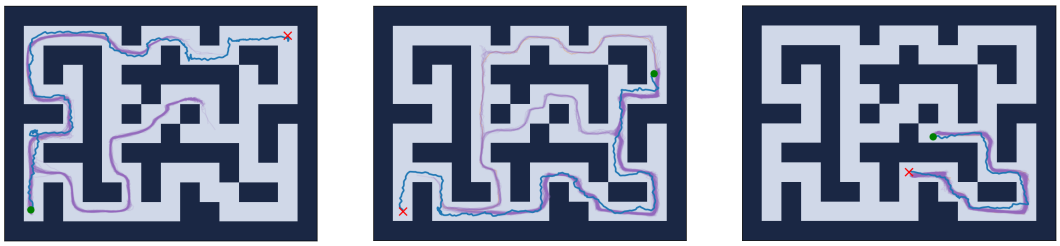
1 INTRODUCTION

Zero-shot RL methods (Touati et al., 2022; Park et al., 2024; Jajoo et al., 2026) based on learning decompositions of the successor measure (SM), have gained prominence in recent years, achieving impressive zero-shot generalization to new tasks in a variety of robot locomotion settings (Tirinzoni et al., 2025; Li et al., 2026). Such models, also known as *behavior foundation models* or BFMs (Pirotta et al., 2024), have the capability to generalize zero-shot to any test-time specified reward function that is representable in the learned feature basis, making them incredibly promising for general-purpose decision-making tasks, such as in the robotics domain. However, these methods have, so far, been limited to relatively short horizon tasks as performance degrades significantly at large discount factors (Azad et al., 2026).

Recent work has instead explored learning a generative model of the successor measure, referred to as a γ-model (Janner et al., 2020) or *geometric horizon model* (GHM) (Thakoor et al., 2022). Such models, trained on top of pretrained BFMs or goal-conditioned policies, can be used for policy improvement (Farebrother et al., 2025; Zheng et al., 2026) and planning (Farebrother et al., 2026) in longer-horizon tasks. However, these methods have many challenges. Firstly, they require significant test-time compute, as each sample from the GHM requires n ODE integration steps, and many samples are necessary to accurately estimate the value function. Moreover, GHMs have so far been trained primarily in raw observation space, making it difficult to extend these methods to larger observation spaces. This is exacerbated by the two stage training process, which first trains a BFM and then learns a GHM on top of the trained policies; The representation space is learned by the BFM and therefore not necessarily compatible with learning accurate GHMs.

It would be preferable to jointly train the BFM and GHM, letting both models influence a shared representation space. However, training these complex flow objectives under shifting policy distributions is a significant obstacle. Recently, Chung et al. (2026) introduced the *universal horizon model* (UHM) in the single-task, offline model-based RL setting. Instead of modeling the full successor measure distribution, UHMs instead condition on time t and learn a policy-dependent time-conditioned world model that directly predicts states t steps ahead. Importantly, this removes a key

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.



(a) Task 1 (b) Task 3 (c) Task 5

Figure 1: Examples of rollouts in `antmaze-giant` from the policy-conditioned UHM versus actual trajectories. 100 rollouts (purple) are generated from the initial state (blue dot) by sampling the UHM recursively for 100 steps, feeding the previous imagined state as the next conditioning state. Each step is sampled from the UHM according to a winsorized geometric distribution with $\beta = 0.9$ and $k_{\text{max}} = 10$. Rollouts are in latent space and decoded using a learned decoder for visualization. We compare against the true policy rollout in the environment (blue). Both UHM and policy are conditioned on the task z, inferred by the reward corresponding to the goal state (red X).

source of stochasticity in GHM predictions in the time step t, such that for approximately deterministic policies and dynamics, a deterministic model of the successor states may suffice. We introduce a novel latent-predictive variation of the UHM model, akin to a JEPA-style world model (Assran et al., 2025). Our model is conditioned on time, as in the UHM, but adopts the latent point-estimate predictions of JEPA world models. In this work, we propose to *jointly* train this latent-predictive UHM alongside TD-JEPA (Bagatella et al., 2026), a recent latent-predictive BFM, with *both* models contributing to shared latent state and task representations. In this context, the UHM acts an auxiliary learning objective for the BFM, which we demonstrate is theoretically principled and experimentally beneficial. We further show that predictions from the UHM can be used to perform model-based updates to augment TD-JEPA training that permits longer horizon BFM learning.

Specifically, our contributions are as follows. We propose a novel policy-conditioned latent-predictive UHM and demonstrate that when trained through bootstrapping in latent space, its fixed point under fixed representations is a low-rank decomposition of an iterated orthogonal projection of the t-step transition matrix. We show that under a Monte Carlo version of the proposed loss, this objective recovers a low-rank decomposition of the t-step transition matrix, and that this decomposition shares representations with the low-rank decomposition of the successor measure under Monte Carlo losses in Bagatella et al. (2026), under certain assumptions. Experimentally, we demonstrate that world model auxiliary losses are modestly beneficial for zero-shot policy learning, leading to an average $\sim 5\%$ increase in success rate across a set of navigation tasks in the OGBench benchmark. Finally, we demonstrate the utility of the learned horizon model by showing that its predictions (see Figure 1) can be used to perform model-based updates, allowing training on discounts as high as $\gamma = 0.999$, achieving performance over double that of model-free TD-JEPA.

2 Related Work

Our work builds on top of zero-shot RL methods (Touati et al., 2022; Bagatella et al., 2026; Park et al., 2024) based on learning a low-rank factorization of the successor measure (Blier et al., 2021). In general, these methods operate by jointly learning a factorization of the policy-conditioned successor measure alongside a class of policies π_z for $z \in \mathcal{Z}$ such that π_z is approximately optimal for a task represented by z. At test-time a dataset of rewards can be used to infer the task vector z to obtain a zero-shot optimal policy for that reward. BYOL-γ (Lawson et al., 2026a) and TD-JEPA (Bagatella et al., 2026) introduce latent-predictive losses for approximating the successor measure in the Monte Carlo and TD cases, respectively. The TD case permits off-policy learning and, hence, can be policy-conditioned for zero-shot RL.

Recent work has explored generative modeling of the successor measure (Farebrother et al., 2025), extending prior work in γ-models (Janner et al., 2020) and geometric horizon models (GHMs) (Thakoor et al., 2022). Farebrother et al. (2025) introduce policy-conditioned GHM modeling of zero-shot RL policies, using a novel TD flow-matching objective to train a class of GHMs condi-

tioned on the same class of policies $\mathcal{Z}$ on which the BFM is trained. This permits fully offline policy improvement through value estimation in the GHM. Farebrother et al. (2026) extend this method to compositional planning over multiple policies. However, this planning requires significant test-time compute. GHMs have also been used for model-based value estimation (MVE) (Janner et al., 2020; Thakoor et al., 2022) and for goal augmentation Lawson et al. (2026b), shifting the computational burden to training time. However, these works only consider a single policy GHM, whereas our work trains a policy-conditioned model.

Besides expensive test-time planning, GHMs also suffer from long-horizon sample degradation (Farebrother et al., 2025). Recently proposed Universal Horizon Models (UHMs) Chung et al. (2026), resolve this limitation by conditioning explicitly on time, allowing for samples to be generated from a broad set of sub-probability measures including bounded distributions such as the winsorized geometric distribution. Importantly, however, UHMs still share significant structure with GHMs: in particular, if the horizon t is sampled from a geometric distribution $\text{Geom}(1-\gamma)$, samples from UHMs correspond directly to those sampled from a GHM with discount γ. Chung et al. (2026)'s UHM is trained for a single policy in raw observation space. In contrast, in this work, we introduce a latent-predictive, policy-conditioned UHM trained through a JEPA-style objective.

Such JEPA-style losses (LeCun et al., 2022), originally introduced in the self-supervised learning setting (Grill et al., 2020) have been explored in RL to learn representations useful for policy learning (Schwarzer et al., 2020). This includes recent work exploring successor measure losses as auxiliary losses for goal-conditioned behavior cloning (Lawson et al., 2026a). Various works have also explored representation learning methods for pretraining features for successor feature learning, including one-step BYOL methods (Jajoo et al., 2026). In contrast, our method explores a multi-step UHM loss, not as a pretraining method, but as an auxiliary loss during BFM training.

3 Background

We consider a reward-free Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \gamma, \rho_0 \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : S \times A \to P(\mathcal{S})$ is a transition function, $\gamma \in [0, 1)$ is the discount factor and $\rho_0 \in P(\mathcal{S})$ is an initial state distribution. A policy π is a mapping $\pi : \mathcal{S} \to P(\mathcal{A})$.

A given policy π in $\mathcal{M}$ emits an unnormalized discounted state visitation distribution, called the *successor measure* (Blier et al., 2021)

$$M^\pi(s, a, \mathcal{X}) = \sum_{t \geq 0} \gamma^t P(s_{t+1} \in \mathcal{X} | s_0 = s, a_0 = a, \pi). \tag{1}$$

The successor measure obeys a forward Bellman equation: $M^\pi = P + \gamma P^\pi M^\pi$, where $P^\pi(s'|s, a) = \sum_a \pi(a|s) P(s'|s, a)$, which inherits the same convergence properties as TD.

For a given feature basis $\psi : \mathcal{S} \to \mathbb{R}^{d_\psi}$, the *successor features* (Barreto et al., 2018) $F^\pi_\psi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\psi}$ are the expected features ψ under M^π

$$F^\pi(s, a) = \int \psi(s') M^\pi(s, a, ds') = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t \psi(s_{t+1}) \mid s_0 = s, a_0 = a, \pi \right]. \tag{2}$$

Importantly, for all rewards linear in the basis ψ, $\mathcal{R}_\psi = \{r(s) = \psi(s)^\top z | z \in \mathbb{R}^{d_\psi}\}$, successor features linearly factorize the Q-function so that $Q^\pi_r(s, a) = F^\pi_\psi(s, a)^\top z_r$. Zero-shot RL methods capitalize on this linear relationship by training *policy-conditioned* successor features $F(s, a, z)$ jointly with a class of policies $\Pi_\mathcal{Z} = \{\pi_z | z \in \mathcal{Z}\}$ where the policies are trained such that $\pi_z(s) \approx \arg\max_a F(s, a, z)^\top z$. That is, the policy π_z is approximately optimal for the reward function defined by the vector z. At test-time, the task is specified by projecting the provided rewards to $\mathcal{Z}$.

3.1 Latent-Predictive Zero-shot RL

BYOL-style latent-predictive losses (Grill et al., 2020) have been applied to representation learning in RL, generally taking the form:

$$\mathcal{L}_{\text{JEPA}}(\phi, T) = \mathbb{E}_{s \sim \rho, s^+ \sim \Xi(\cdot|s)} ||T(\phi(s)) - \overline{\phi(s^+)}||^2 \tag{3}$$

where $\phi(s) : \mathcal{S} \to \mathbb{R}^{d_\phi}$ is a state encoder, $T : \mathbb{R}^{d_\phi} \to \mathbb{R}^{d_\phi}$ is a latent-space predictor, and Ξ is some distribution over states, such as the transition kernel P^π (Schwarzer et al., 2020) or the successor measure M^π (Lawson et al., 2026a).

In a recent work, Bagatella et al. (2026) propose a latent-predictive zero-shot RL method which learns policy-conditioned successor features $T_\phi(\phi(s), a, z) = F_\psi(\phi(s), a, z)$ over task basis ψ and state-encoder ϕ through the latent predictive loss

$$\mathcal{L}_{\text{MC-JEPA}}(\phi, T_\phi, \psi) = \mathbb{E}_{\substack{(s,a)\sim\mathcal{D}, z\sim\mathcal{Z}, \\ s^+\sim M^{\pi_z}(\cdot|s,a)}} \left[||T_\phi(\phi(s), a, z) - \overline{\psi(s^+)}||^2 \right]. \tag{4}$$

This can be learned from off-policy data via minimizing the Bellman loss over successor features:

$$\mathcal{L}_{\text{TD-JEPA}}(\phi, T_\phi, \psi) = \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D}, z\sim\mathcal{Z}, \\ a'\sim\pi_z(\cdot|s')}} \left[||T_\phi(\phi(s), a, z) - \overline{\psi(s')} - \gamma \overline{T_\phi(\phi(s'), a', z)}||^2 \right]. \tag{5}$$

Several works have studied the theoretical properties of these latent predictive losses (Tang et al., 2023; Khetarpal et al., 2024; Lawson et al., 2026a). Bagatella et al. (2026) generalize these results by demonstrating that for the general case of Equation (4), for any sampling density over s^+, the latent-predictive loss descends towards the best multilinear factorization of the density kernel.

3.2 Geometric and Universal Horizon Models

Geometric Horizon Models (GHMs) are generative models that directly model the normalized successor measure in Equation (1) (Janner et al., 2020; Thakoor et al., 2022; Farebrother et al., 2025). In contrast, a Universal Horizon Model (UHM), as introduced in Chung et al. (2026), is a *time-conditioned* generative model of the transition measure under a policy π: given a horizon $t \in \mathbb{N}$, the UHM $m^\pi(x|s, a, t) = \Pr(s_t = x|s_0 = s, a_0 = a, \pi)$ models the distribution over states that are t steps away from s when following π.

This generalizes Geometric Horizon Models (GHMs) by allowing sampling from arbitrary horizon distributions $p_h(t) \in \Delta(\mathbb{N})$, with $p_h(t) = \text{Geom}(1-\gamma)$ equivalent to sampling from a γ-GHM.

In Chung et al. (2026), the authors train the UHM for a single policy π using recursion:

$$m^\pi(x|s, a, t) = \begin{cases} P(s'|s, a) & \text{if } t = 1 \\ \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')} \left[m^\pi(x|s', a', t-1) \right] & \text{if } t > 1. \end{cases} \tag{6}$$

They show that for any sub-probability measure $\nu(t)$ over $\mathbb{N}$ (i.e. $\nu(k) \geq 0, \sum_{k\geq 1} \nu(k) \leq 1$), the following ν-Bellman operator converges to $Q^\pi(s, a)$:

$$\mathcal{T}^\nu Q(s, a) := \mathbb{E}\left[r(s, a) + \gamma \left[\sum_{k\geq 1} \xi(k) r(s_k, a_k) + \sum_{k\geq 1} \nu(k) Q(s_k, a_k) \right] \right] \tag{7}$$

where

$$\xi(k) = \gamma^{k-1} - \sum_{i=1}^{k} \gamma^{k-i} \nu(i). \tag{8}$$

Estimating the operator in Equation (7) can then be done with model-based samples:

$$G_t = r(s, a) + \gamma \mathbb{E}_{k\sim p_h, s^+\sim m^\pi(\cdot|s,a,k), a^+\sim\pi(\cdot|s^+)} \left[w_\xi(k) r(s^+, a^+) + w_\nu(k) \overline{Q(s^+, a^+)} \right] \tag{9}$$

where $w_\xi(k) = \frac{\xi(k)}{p_h(k)}$, $w_\nu(k) = \frac{\nu(k)}{p_h(k)}$ are importance weights.

4 Method

Our proposed method is an extension to the TD-JEPA algorithm (see Section 3.1), which additionally (a) trains a latent-predictive UHM and (b) augments standard TD-JEPA training with MVE-style model-based updates using samples from the UHM. We discuss each component in detail in the following sections. For simplicity, we first describe our method for the symmetric case where $\phi = \psi$; however, we provide theoretical analysis for the full asymmetric case. Following Bagatella et al. (2026), for theoretical results, we will consider the action-free, tabular setting where predictors are linear. The full algorithm is provided in Algorithm 1.

4.1 LATENT-PREDICTIVE UNIVERSAL HORIZON MODEL

Concurrently with the SF model, we train a latent-predictive UHM over the same base features ϕ (and ψ in the asymmetric case). Both models are trained jointly, and the representations are shaped by both the UHM and SF losses. Hence, our method adds only a separate UHM predictor, which we denote by $T_\phi(\phi(s), a, z, t)$, where t conditioning differentiates from the unconditioned TD-JEPA successor features predictor $T_\phi(\phi(s), a, z)$.

Adapting Equation (6) from Chung et al. (2026) to the latent-predictive setting, we train our UHM with the bootstrapped loss

$$\mathcal{L}_{\text{UHM-JEPA}}(\phi, T_\phi, t) = \begin{cases} \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D}, \\ z\sim\mathcal{Z}}} ||T_\phi(\phi(s), a, z, t) - \overline{\phi(s')}||^2 & \text{if } t = 1 \\ \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D}, \\ z\sim\mathcal{Z}, \\ a'\sim\pi_z(\cdot|s)}} ||T_\phi(\phi(s), a, z, t) - \overline{T_\phi(\phi(s'), a', z, t-1))}||^2 & \text{if } t > 1 \end{cases}, \tag{10}$$

where t can be sampled from any discrete probability distribution over the finite horizon $\{1, ..., H\}$ (see Section A.1).

Thus, for $t = 1$, the model is trained on the latent representation of the next state from the dataset and for $t > 1$ the model is trained to be self-consistent in predictions conditioned on both t and $t-1$, respectively. Notably, unlike in Chung et al. (2026) our world model is policy-conditioned through z, in the same style as zero-shot RL methods (Touati et al., 2022; Bagatella et al., 2026) using the jointly trained policy π_z to sample the action a' at next state s', allowing us to learn a family of UHMs for the same policy family $\mathcal{Z}$ on which we train successor features.

Unlike in Chung et al. (2026), which trains the UHM on raw state space, our method proposes to jointly learn latent representations through gradients from both the UHM-JEPA and TD-JEPA losses (Equation (5)). We demonstrate that for Monte Carlo version of the two losses, **MC-UHM-JEPA** and **MC-JEPA** (Equation (4)), respectively, the representations learned are compatible in the sense that for a fixed z, each loss descends towards a d-rank factorization of their respective kernels, and these kernels share top-d eigenvectors, assuming the top-d eigenvalues are positive and distinct.

We consider an asymmetric Monte Carlo version of the UHM loss from Equation (16):

$$\mathcal{L}_{\text{MC-UHM-JEPA}}(\phi, T_\phi, t, \psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}, z\sim\mathcal{Z}, s^+\sim(P^{\pi_z})^t(s,a)} ||T_\phi(\phi(s), a, z, t) - \overline{\psi(s^+)}||^2. \tag{11}$$

Under the same assumptions as prior work (Khetarpal et al., 2024; Lawson et al., 2026a; Bagatella et al., 2026), namely a uniform state distribution (A1), identity covariances (A2) and symmetric transitions (A3), we can show that the tabular **MC-UHM-JEPA** loss corresponds to Equations 22 and 23 in Proposition 2 of Bagatella et al. (2026) for density $(P^{\pi_z})^t$ (see Section A.1). Therefore, by direct application of Theorem 5 in Bagatella et al. (2026), we have that the density loss

$$\mathcal{L}_{\text{dens}}^{(P^{\pi_z})^t}(\phi, T_{z,t}, \psi, t) = \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z}} ||\phi T_{z,t} \psi^\top - (P^{\pi_z})^t||_F^2 \tag{12}$$

is a Lyapunov function for the ODE induced by optimizing tabular **MC-UHM-JEPA** losses under the usual two-time scale assumptions (Tang et al., 2023) and hence optimizing the **MC-UHM-JEPA** losses descends towards the best multi-linear factorization of the t-step transition matrices.

Similarly, as is shown in Bagatella et al. (2026), the density loss

$$\mathcal{L}_{\text{dens}}^{M^{\pi_z}}(\phi^M, T_z^M, \psi^M) = \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z}} ||\phi^M T_z^M \psi^{M\top} - M^{\pi_z}||_F^2 \tag{13}$$

is a Lyapunov function of the **MC-JEPA** losses and hence optimizing the **MC-JEPA** losses descends towards the best multi-linear factorization of the successor measure matrix.

From here, we note that P^{π_z} and M^{π_z} share eigenvectors since $M = (I - \gamma P)^{-1}$ (Chandak et al., 2023). The Frobenius norms in Equation (29) and Equation (32) find the best d-rank approximations to $(P^{\pi_z})^t$ and M^{π_z} by selecting the singular vectors corresponding to the d largest singular values. Under assumption A3, P^{π_z} (and hence M^{π_z}) are symmetric so their singular vectors are the eigenvectors and singular values are given by $|\lambda|^t$ and $|\frac{1}{1-\gamma\lambda}| = \frac{1}{1-\gamma\lambda}$, respectively. For any $\lambda \geq 0$, these orderings are consistent, leading to Theorem 1.

Theorem 1 (Compatible Representations under Monte Carlo Losses) *Considering the following loss:*

$$\mathcal{L}_{\text{MC-UHM-JEPA}}(\phi, T_{\phi,z,t}, \psi) = \mathbb{E}_{\substack{t\sim p_h, s\sim\mathcal{D},\\ z\sim\mathcal{Z}, s^+\sim(P^{\pi_z})^t(\cdot|s)}} ||T^\top_{\phi,t,z}\phi(s) - \overline{\psi(s^+)}||^2. \tag{14}$$

Let (ϕ^*, ψ^*) be the joint minimizers of the losses $\mathcal{L}_{\text{MC-UHM-JEPA}}(\phi, T_{\phi,z,t}, \psi)$ and $\mathcal{L}_{\text{MC-UHM-JEPA}}(\psi, T_{\psi,z,t}, \phi)$, and let $(\phi^{M,*}, \psi^{M,*})$ be the joint minimzers of the losses $\mathcal{L}_{\text{MC-JEPA}}(\phi, T_{\phi,z}, \psi)$ and $\mathcal{L}_{\text{MC-JEPA}}(\psi, T_{\psi,z}, \phi)$. Assuming A1, A2, A3 and that the eigenvalues of P^{π_z} for all $z \in \mathcal{Z}$ are positive and distinct, for $|\mathcal{Z}| = 1$ (single-policy setting), we have:

$$span(\phi^*) = span(\phi^{M,*}), \quad span(\psi^*) = span(\psi^{M,*}) \tag{15}$$

where span indicates the column span of the matrices. For $|\mathcal{Z}| > 1$, the same holds when P^{π_z} share a top-d subspace $\forall z \in \mathcal{Z}$.

Proof. See Section A.1

Hence, both losses share a common set of minimizing solutions.

However, in off-policy settings, $(P^{\pi_z})^t$ and M^{π_z} are not available to sample from. In the case of M^{π_z}, Bagatella et al. (2026) extend the standard TD equations for the successor measure to derive a TD bootstrap approximation to the factorization. Similarly, the UHM setting emits a recursive bootstrap for the t-step transition matrix (Chung et al., 2026) , which we extend to the latent-predictive setting, leading to the bootstrapped loss proposed in Equation (16). This bootstrapped UHM loss **UHM-JEPA** instead learns a low rank factorization of an *iterated orthogonal projection* of the t-step transition matrix.

Theorem 2 (Fixed Point of Latent-Predictive UHM) *Consider the following loss:*

$$\mathcal{L}_{\text{UHM-JEPA}}(\phi, T_{\phi,z,t}, \psi) = \begin{cases} \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D},\\ z\sim\mathcal{Z}}} ||T^\top_{\phi,z,t}\phi(s) - \overline{\psi(s')}||_2^2 & \text{if } t = 1 \\ \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D},\\ z\sim\mathcal{Z},\\ a'\sim\pi_z(\cdot|s)}} ||T^\top_{\phi,z,t}\phi(s) - \overline{T^\top_{\phi,z,t-1}\phi(s')}||^2 & \text{if } t > 1 \end{cases} \tag{16}$$

For fixed ϕ, ψ let $T^*_{\phi,t,z}$ and $T^*_{\psi,t,z}$ be the optimal predictors for $\mathcal{L}_{\text{UHM-JEPA}}(\phi, T_{\phi,z,t}, \psi)$ and $\mathcal{L}_{\text{UHM-JEPA}}(\psi, T_{\psi,z,t}, \phi)$, respectively. Then, under the same assumptions A1, A2, A3, for all z, t, $\phi T^*_{\phi,z,t} = (\Pi_\phi P^{\pi_z})^t\psi$ and $\psi T^*_{\psi,z,t} = (\Pi_\psi P^{\pi_z})^t\phi$ where Π_ϕ and Π_ψ are orthogonal projections onto the span of ϕ and ψ, respectively.

Proof. See Section A.2.

4.2 Model-Based TD-JEPA Updates

The second component of our algorithm utilizes the latent-predictive UHM to perform model-based TD-JEPA updates of the form

$$\mathcal{L}_{\text{MB-TD-JEPA}} = \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D}, z\sim\mathcal{Z}, k\sim p_h(\cdot),\\ \phi(s_k^+)=T_{\phi,z}(\phi(s'),a',z,k),\\ a^+\sim\pi_z(\phi(s_k^+))}} ||T_{\phi,z}(\phi(s), a, z) - \overline{\phi(s')} - \gamma\left(w_\xi(k)\overline{\phi(s_k^+)} + w_\nu(k)\overline{T_{\phi,z}(\phi(s_k^+), a^+, z)}\right)||^2 \tag{17}$$

where $w_\xi = \frac{\xi(k)}{p_h(k)}$ and $w_\nu = \frac{\nu(k)}{p_h(k)}$, $\nu(k)$ a sub-probability measure, and ξ_ν defined in Equation (19).

This follows from an extension of the ν-Bellman equation (Equation (7)) to the successor measure.

Theorem 3 (ν-Bellman Operator on Successor Measures) *The successor measure with discount γ under a policy π, denoted M^π_γ, obeys the ν-Bellman equation*[1]

$$M^\pi_\gamma = P + \gamma\sum_{k\geq 1}\xi(k)(P^\pi)^{k+1} + \gamma\sum_{k\geq 1}\nu(k)(P^\pi)^{k+1}M^\pi_\gamma \tag{18}$$

[1]The index shift by one in Equation (19) versus Equation (8) is due to factoring out the initial transition

where

$$\xi(k) = \gamma^{k-1} - \sum_{i=1}^{k-1} \nu(i)\gamma^{k-i-1} \quad \forall \quad k \geq 1 \tag{19}$$

for any sub-probability measure ν. Furthermore, the Bellman operator given in Equation 18 is a contraction when $|\gamma| < 1$ for any ν.

Proof. See Section A.3.

Intuitively, this update splits the bootstrap term from the standard TD-JEPA loss into two steps: a jump to $\phi(s_k^+)$ from $\phi(s')$ according to the horizon distribution $\nu(k)$, with importance sampling weights reweighting samples drawn from p_h, and a final step from $\phi(s_k^+)$ onward. In this way, it is a generalization of the horizon-consistency objective from Farebrother et al. (2026) to sub-probability measures over k. In Section A.5, we show that we can recover the horizon consistency loss from Farebrother et al. (2026) when $\nu(k) = (\gamma - \beta)\beta^{k-1}$.

In our latent-predictive formulation, we characterize the fixed point of the predictor under fixed representations as a factorization of an oblique projection of the successor measure onto the span of ϕ. This is consistent with the fixed point results for standard TD-JEPA which also learns a factorization of an oblique projection of the successor measure (Bagatella et al., 2026).

Theorem 4 (Fixed Point of Model-Based TD-JEPA) *Consider the following loss*

$$\mathcal{L}_{\textit{MB-TD-JEPA}}(\phi, T_{\phi,z}, \psi) \tag{20}$$

$$= \mathbb{E}_{\substack{s\sim\rho, z\sim\mathcal{Z}, \\ s'\sim P^{\pi_z}(\cdot|s), \\ k\sim p_h, s^+\sim(P^{\pi_z})^k(\cdot|s')}} ||T_{\phi,z}^\top \phi(s) - \overline{\psi(s')} - \gamma(w_\xi(k)\overline{\psi(s^+)} + w_\nu(k)\overline{T_{\phi,z}^\top \phi(s^+)})||_2^2 \tag{21}$$

For fixed ϕ, ψ let $T^*_{\phi,z}$ and $T^*_{\psi,z}$ be the semi-gradient TD optimal predictors for $\mathcal{L}_{\textit{MB-TD-JEPA}}(\phi, T_{\phi,z}, \psi)$ and $\mathcal{L}_{\textit{MB-TD-JEPA}}(\psi, T_{\psi,z}, \phi)$, respectively. Then, under the same assumptions A1, A2, A3, for all z, $\phi T^*_{\phi,z} = \tilde{\Pi}_\phi M^{\pi_z}\psi$ and $\psi T^*_{\psi,z} = \tilde{\Pi}_\psi M^{\pi_z}\phi$ where $\tilde{\Pi}_\phi$ and $\tilde{\Pi}_\psi$ are oblique projections onto the span of ϕ and ψ, respectively.

5 Experiments

In our experiments, we validate that latent-predictive zero-shot RL can successfully train at higher γ through model-based updates using predictions from a jointly trained UHM. To test this, we train both standard TD-JEPA and our proposed model-based TD-JEPA (denoted **MB-TD-JEPA**) at varying γ on the OGBench navigation environments (Park et al., 2025a) used in Bagatella et al. (2026), with the addition of the `antmaze-giant-navigate` environment which has a longer task horizon. These benchmarks span different environment sizes and dataset compositions. Additionally, we consider the effect of the UHM auxiliary loss in isolation (denoted **+UHM**) .

In order to adapt TD-JEPA for the the model-based setting, we make several changes to the implementation as reported in Bagatella et al. (2026). First, to simplify the learning process, we use the symmetric variant of TD-JEPA and MB-TD-JEPA. Second, since our method assumes roughly deterministic policies, we use a plain BC penalty on the actor instead of the FlowQ-like regularization (Park et al., 2025b) used in TD-JEPA. Third, our method requires bootstrapping off imagined states from the UHM, and hence it is necessary to condition the policy on the state embedding $\phi(s)$ instead of the raw state s as is done in Bagatella et al. (2026). We note that this change is theoretically consistent with the TD-JEPA algorithm, since the Q-values that the policy optimizes are only conditioned on the latent state and therefore the policy should not require additional state information besides that contained in the latent state. Finally, we learn a normalized version of the successor measure, i.e. $\tilde{M} = (1-\gamma)M$. This ensures that the TD-JEPA loss and the UHM loss are on consistent scales. We report TD-JEPA performance for the exactly matched setup to our experiments (denoted **TD-JEPA (Matched)**). For context, we additionally provide the original TD-JEPA results (denoted **TD-JEPA (Original)**) as reported in Bagatella et al. (2026) in Table 2.

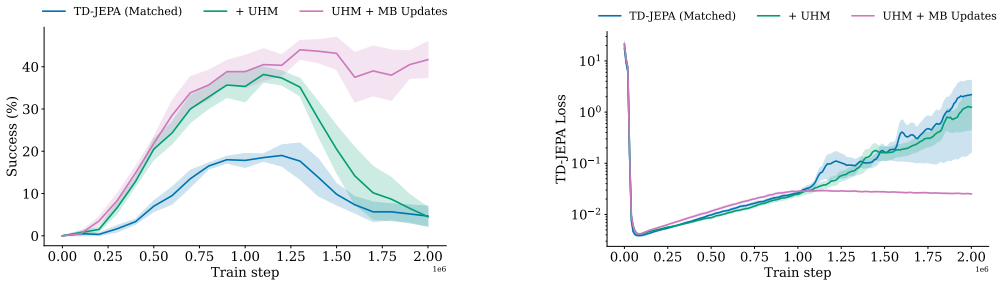
In Figure 1 we provide a qualitative evaluation of the quality of the UHM predictions, where we see that our latent-predictive UHM can faithfully produce task-conditioned trajectories that mimic the true policy rollouts. Thus, we might expect these predictions to be useful for model-based updates.

Table 1: Average final success rate across tasks on OGBench `antmaze` environments at $\gamma = 0.995$ and $\gamma = 0.999$. Average success is computed as the average of the last three evaluations during training where each evaluation consists of 10 episodes per each of five tasks. Mean and standard error are reported across 10 seeds. Per environment and γ, the best-performing variant is bolded, along with any variant whose mean plus one standard error reaches the best mean.

	TD-JEPA (Matched)	+ UHM	MB-TD-JEPA
	$\gamma = 0.995$		
antmaze-medium-stitch	52.27 ± 3.22	**55.93 ± 11.37**	**60.80 ± 3.12**
antmaze-medium-explore	19.40 ± 4.20	**40.07 ± 4.73**	33.07 ± 2.21
antmaze-medium-navigate	**92.13 ± 0.99**	**92.67 ± 1.11**	**92.00 ± 0.86**
antmaze-large-stitch	19.07 ± 1.54	**24.53 ± 1.64**	20.53 ± 1.33
antmaze-large-navigate	77.93 ± 1.14	76.93 ± 1.23	**81.20 ± 0.97**
antmaze-giant-navigate	34.60 ± 2.32	**43.33 ± 1.36**	39.24 ± 1.83
Mean	49.23	**55.58**	54.47
	$\gamma = 0.999$		
antmaze-medium-stitch	27.40 ± 5.41	33.20 ± 5.17	**41.33 ± 3.34**
antmaze-medium-explore	0.13 ± 0.09	0.47 ± 0.14	**22.27 ± 2.94**
antmaze-medium-navigate	45.40 ± 5.12	56.33 ± 6.43	**91.47 ± 0.38**
antmaze-large-stitch	5.07 ± 1.02	6.87 ± 0.76	**9.20 ± 1.75**
antmaze-large-navigate	24.60 ± 3.75	36.53 ± 4.40	**67.47 ± 4.11**
antmaze-giant-navigate	4.07 ± 0.49	3.00 ± 0.36	**39.67 ± 1.77**
Mean	17.78	22.73	**45.34**

In Table 1 we provide the results for training with standard TD-JEPA, with auxiliary UHM loss and with full MB-TD-JEPA over both $\gamma = 0.995$ and $\gamma = 0.999$. Firstly, we find that at both discounts the auxiliary UHM loss alone provides some benefit over standard TD-JEPA learning, with five of the six environments showing improvement over TD-JEPA. However, at $\gamma = 0.995$, the additional effect of model-based updates is neutral, with some environments showing improvement and others matching or slightly worse than with the UHM auxiliary loss alone.

The story is quite different in the high discount setting, however. Here, performance collapses in several environments for both standard TD-JEPA and the UHM auxiliary loss variant. Across all six environments, MB-TD-JEPA outperforms both baselines by a considerable margin, achieving average performance of 45%, nearly double the average with UHM auxiliary loss alone.



(a) Success Rate (b) TD-JEPA Loss

Figure 2: Learning curves (a) success rate and (b) TD loss for TD-JEPA and MB-TD-JEPA in the `antmaze-giant` environment. For each method, we select the learning curve for the hyperparameters with the best *ever* performance during training instead of best final performance used in Table 1. This is to better illustrate the catastrophic failure mode exhibited during learning without model-based updates. Curves average over four seeds with standard error shading.

To better understand the mechanism by which MB-TD-JEPA outperforms in the high discount setting, we provide an example of learning curves for `antmaze-giant` in Figure 2 (see Section C for all environments). Here, we can see a clear pattern that demonstrates the value of model-based updates at high discount rates. TD-JEPA, both with and without the UHM auxiliary loss, diverges in late training with losses exploding after one million training steps (Figure 2b). This divergence

leads to a catastrophic collapse in performance (Figure 2a). In contrast, with model-based updates, training loss is stable and policy performance remains consistent through late training.

Next, we consider the interesting result that the auxiliary UHM loss boosts performance across the `antmaze` environments. In Table 2, we report the performance of the UHM auxiliary loss and standard TD-JEPA at $\gamma = 0.995$ along with two ablations: (1) a simple one-step loss which is equivalent to the UHM loss where k is sampled from the Dirac delta on 1 and (2) a version of our loss where a one-step auxiliary loss is trained in addition to the UHM loss. Note that the UHM loss itself puts some weight on the one-step transition loss, according to the k sampling distribution; the auxiliary one-step loss is an additional loss term computed over the entire batch.

Table 2: Average success rate across tasks on OGBench `antmaze` environments at $\gamma = 0.995$. Average success is computed as the average of the last three evaluations during training. Mean and standard error are reported across 10 seeds. TD-JEPA (Original) results are reproduced from Bagatella et al. (2026). Per environment, the best-performing variant is bolded, along with any variant whose mean plus one standard error reaches the best mean (excluding TD-JEPA (Original))

.

	TD-JEPA (Original)	TD-JEPA (Matched)	+ 1-Step	+ UHM	+ UHM w/ 1-Step
antmaze-medium-stitch	61.56 ± 4.53	52.27 ± 3.22	**53.27 ± 6.42**	**55.93 ± 11.37**	**56.80 ± 2.53**
antmaze-medium-explore	20.20 ± 2.39	19.40 ± 4.20	13.60 ± 2.17	**40.07 ± 4.73**	19.07 ± 3.64
antmaze-medium-navigate	70.40 ± 3.72	92.13 ± 0.99	**95.93 ± 0.59**	92.67 ± 1.11	93.40 ± 0.73
antmaze-large-stitch	40.60 ± 2.51	19.07 ± 1.54	**34.67 ± 2.55**	24.53 ± 1.64	21.13 ± 1.54
antmaze-large-navigate	57.20 ± 4.25	77.93 ± 1.14	**86.80 ± 1.27**	76.93 ± 1.23	**85.80 ± 1.67**
antmaze-giant-navigate	—	34.60 ± 2.32	45.93 ± 2.17	43.33 ± 1.36	**52.20 ± 3.04**
Mean		49.23	55.03	**55.58**	54.73

We find that the benefit of training with an auxiliary world model loss is not exclusive to UHM training, with all world-model auxiliary losses leading on average to a modest boost in performance of $\sim 5\%$ across the `antmaze` environments. We hypothesize that this may be due to the world model losses facilitating a better selection of top-d singular vectors in the decomposition of the successor measure. While we show in Section 4.1 that for a positive top-d spectrum, under certain conditions, the decompositions learned by the UHM and TD-JEPA should align; in practice, the magnitude of the eigenvalues can change the learning dynamics. As noted in prior work (Touati et al., 2022), increasing γ sharpens the spectra of the successor measure matrix, accentuating differences between large-magnitude singular values, which facilitates training at larger horizons while hampering it at shorter ones. Jointly training the UHM with $\beta < \gamma$ may thus provide better learning signal for singular vectors with lower-magnitude singular values.

6 Conclusions, Limitations and Future Work

In this work, we establish a novel latent-predictive model-based zero-shot RL method, and validate the hypothesis that model-based methods can stabilize learning at higher values of γ. Further, we find that training world model auxiliary losses, either one-step or multi-step, concurrently with standard TD-JEPA is modestly beneficial to downstream zero-shot policy learning.

As our focus is on establishing the *validity* of model-based updates for stabilizing high γ training, our evaluation and conclusions are limited to proprioceptive navigation environments in OGBench. Manipulation environments present additional challenges, currently requiring conditioning the policy on raw states (Bagatella et al., 2026), which introduces complexity that we leave to future work. Though we believe our method is well-suited to pixel-based environments, as it operates in latent space, we also leave this as a direction for future work.

Additionally, we acknowledge that the assumptions of Theorem 1 are restrictive. The assumption of positive spectra may be violated in environments which have strong periodicity, for example. Though we believe the assumption of a shared top-d subspace of P^{π_z} across $z \in \mathcal{Z}$ could potentially be relaxed under certain horizon sampling distributions, we leave this exploration to future work.

Finally, the additional auxiliary loss and model-based updates expand the hyperparameter search space for tuning. Since training divergence at high γ can be evaluated entirely offline, we believe this could offer a signal for choosing hyperparameters without online evaluations, though we leave this exploration to future work.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods via coding assistance and generation, and to help with formulating and verifying mathematical claims and proofs. All LLM-generated code was reviewed and verified by the authors. We have not used generative AI tools to generate synthetic data sets, provide critical ingredients for proving mathematical claims, propose hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis or interpret results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide extensive details of our methodology including, architecture, training details and hyperparameters in Section B. Our code is built on a public code base (link) and our code will be made public upon publication.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Seyed Mahdi B. Azad, Jasper Hoffmann, Iman Nematollahi, Hao Zhu, Abhinav Valada, and Joschka Bödecker. Spectral alignment in forward-backward representations via temporal abstraction, 2026.

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

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International conference on machine learning*, pp. 501–510. PMLR, 2018.

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

Yash Chandak, Shantanu Thakoor, Zhaohan Daniel Guo, Yunhao Tang, Remi Munos, Will Dabney, and Diana L Borsa. Representations and exploration for deep reinforcement learning using singular value decomposition. In *International Conference on Machine Learning*, pp. 4009–4034. PMLR, 2023.

Hojun Chung, Junseo Lee, and Songhwai Oh. Offline reinforcement learning with universal horizon models. *arXiv preprint arXiv:2605.15603*, 2026.

Jesse Farebrother, Matteo Pirotta, Andrea Tirinzoni, Remi Munos, Alessandro Lazaric, and Ahmed Touati. Temporal difference flows. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16102–16143. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/farebrother25a.html.

Jesse Farebrother, Matteo Pirotta, Andrea Tirinzoni, Marc G Bellemare, Alessandro Lazaric, and Ahmed Touati. Compositional planning with jumpy world models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cwztmjMQ77.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar,

et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Pranaya Jajoo, Harshit Sikchi, Siddhant Agarwal, Amy Zhang, Scott Niekum, and Martha White. Regularized latent dynamics prediction is a strong baseline for behavioral foundation models. *arXiv preprint arXiv:2603.15857*, 2026.

Michael Janner, Igor Mordatch, and Sergey Levine. γ-models: Generative temporal difference learning for infinite-horizon prediction. In *Advances in Neural Information Processing Systems*, 2020.

Khimya Khetarpal, Zhaohan Daniel Guo, Bernardo Avila Pires, Yunhao Tang, Clare Lyle, Mark Rowland, Nicolas Heess, Diana Borsa, Arthur Guez, and Will Dabney. A unifying framework for action-conditional self-predictive reinforcement learning. *arXiv preprint arXiv:2406.02035*, 2024.

Daniel Lawson, Adriana Hugessen, Charlotte Cloutier, Glen Berseth, and Khimya Khetarpal. Self-predictive representations for combinatorial generalization in behavioral cloning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=FkeURAdA0h.

Daniel Lawson, Dhruv Sreenivas, Adriana Hugessen, Khimya Khetarpal, and Glen Berseth. Imagining goals for policy generalization with geometric horizon models. In *Third Reinforcement Learning Beyond Rewards Workshop at the Reinforcement Learning Conference*, 2026b. URL https://openreview.net/forum?id=rL2Jrd7SRX.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, Alessandro Lazaric, Matteo Pirotta, and Guanya Shi. BFM-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jkhl2oI0g5.

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

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=M992mjgKzI.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. *arXiv preprint arXiv:2502.02538*, 2025b.

Matteo Pirotta, Andrea Tirinzoni, Ahmed Touati, Alessandro Lazaric, and Yann Ollivier. Fast imitation via behavior foundation models. In *International Conference on Learning Representations*, volume 2024, pp. 12685–12724, 2024.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

Yunhao Tang, Zhaohan Daniel Guo, Pierre Harvey Richemond, Bernardo Avila Pires, Yash Chandak, Rémi Munos, Mark Rowland, Mohammad Gheshlaghi Azar, Charline Le Lan, Clare Lyle, et al. Understanding self-predictive learning for reinforcement learning. In *International conference on machine learning*, pp. 33632–33656. PMLR, 2023.

Shantanu Thakoor, Mark Rowland, Diana Borsa, Will Dabney, Remi Munos, and Andre Barreto. Generalised policy improvement with geometric policy composition. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21272–21307. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/thakoor22a.html.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9sOR0nYLtz.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? *arXiv preprint arXiv:2209.14935*, 2022.

Chongyi Zheng, Seohong Park, Sergey Levine, and Benjamin Eysenbach. Intention-conditioned flow occupancy models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j6D83Mf6LG.

A Theoretical Results

In the following results, we assume an action-free, finite MDP with linear predictors.

Let $\mathcal{Z}$ be finite, and for each $z \in \mathcal{Z}$ let $\Xi_z \in \mathbb{R}^{S\times S}$ be a generic kernel. Let D_ρ be the diagonal matrix of the state distribution ρ, write $\Sigma_\phi := \phi^\top D_\rho \phi$, $\Sigma_\psi := \psi^\top D_\rho \psi$, and let $\Xi_z^* := D_\rho^{-1}\Xi_z^\top D_\rho$ be the ρ-adjoint of Ξ_z.

We will refer to the following three assumptions:

- A1 State distribution is uniform: $D_\rho = I$
- A2 Identity covariances: $\Sigma_\phi = \Sigma_\psi = I$
- A3 Symmetric dynamics: $P^{\pi_z} = {P^{\pi_z}}^\top \quad \forall z \in \mathcal{Z}$

A.1 Proof of Theorem 1

First, we briefly restate Theorem 5 and Proposition 4 from Bagatella et al. (2026):

Define the density-based loss

$$L_{\text{dens}}(\phi, T_z, \psi) := \frac{1}{2}\,\mathbb{E}_{z\sim\mathcal{Z}}\Big[\big\|D_\rho^{1/2}(\phi T_z \psi^\top - \Xi_z D_\rho^{-1})D_\rho^{1/2}\big\|_F^2\Big], \tag{22}$$

and the latent-predictive losses

$$L_\phi(\phi, T_z, \psi) := \tfrac{1}{2}\,\mathbb{E}_{z\sim\mathcal{Z}}\big[\|D_\rho^{1/2}(\phi T_z - \overline{\Xi_z \psi \Sigma_\psi^{-1}})\|_{\Sigma_\psi}^2\big], \tag{23}$$

$$L_\psi(\psi, T_z, \phi) := \tfrac{1}{2}\,\mathbb{E}_{z\sim\mathcal{Z}}\big[\|D_\rho^{1/2}(\psi T_z - \overline{\Xi_z^* \phi \Sigma_\phi^{-1}})\|_{\Sigma_\phi}^2\big], \tag{24}$$

where $\|X\|_W := \|XW^{1/2}\|_F$.

Theorem 5 (Bagatella et al., 2026, Thm. 5 & Prop. 2) *For any kernel Ξ_z, the latent-predictive losses L_ϕ, L_ψ in equation 23–equation 24 satisfy the gradient-matching conditions for L_{dens}; consequently L_{dens} is a Lyapunov function for the ODE that dictates the optimization dynamics when T_z is optimized to convergence at each iteration.*

We will show that this Theorem directly applies to our **MC-UHM-JEPA** loss. Let the horizon sampling distribution $\mathcal{T}$ be finite. We write Equation (11) in matrix form and take the expectation over $\mathcal{T}$

$$\mathcal{L}_{\text{MC-UHM-JEPA}}(\phi, T_{z,t}, \psi) = \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z},t\sim\mathcal{T}}||D_\rho^{1/2}(\phi T_{z,t} - \overline{(P^{\pi_z})^t \psi})||_F^2 \tag{25}$$

.

Let $\mathcal{Z}' : \mathcal{T} \times \mathcal{Z}$ be the joint distribution over tasks and horizons, which is also finite. Then, under assumption A1 we have

$$\mathcal{L}_{\text{MC-UHM-JEPA}}(\phi, T_{z,t}, \psi) = \frac{1}{2}\mathbb{E}_{(z,t)\sim\mathcal{Z}'}||(\phi T_{z,t} - \overline{(P^{\pi_z})^t \psi})||_F^2 \tag{26}$$

,

while under assumptions A1, A2, and A3, and pluggin in $\mathcal{Z}'$ for $\mathcal{Z}$, we have

$$L_\phi(\phi, T_{z,t}, \psi) := \tfrac{1}{2}\,\mathbb{E}_{(z,t)\sim\mathcal{Z}'}\left[\|(\phi T_{z,t} - \overline{\Xi_{z,t}\psi})\|_F^2\right], \tag{27}$$

$$L_\psi(\psi, T_{z,t}, \phi) := \tfrac{1}{2}\,\mathbb{E}_{(z,t)\sim\mathcal{Z}'}\left[\|(\psi T_{z,t} - \overline{\Xi_{z,t}\phi})\|_F^2\right], \tag{28}$$

Clearly, these are equivalent for $\Xi_{z,t} = (P^\pi)^t$. Therefore, by direct application of Theorem 5 in Bagatella et al. (2026), we have that the density loss

$$\mathcal{L}_{\text{dens}}^{(P^{\pi_z})^t}(\phi, T_{z,t}, \psi) = \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z},t\sim\mathcal{T}}||\phi T_{z,t}\psi^\top - (P^{\pi_z})^t||_F^2 \tag{29}$$

is a Lyapunov function for the ODE induced by optimizing losses $\mathcal{L}_{\text{MC-UHM-JEPA}}(\phi, T_\phi, t, \psi)$ and $\mathcal{L}_{\text{MC-UHM-JEPA}}(\psi, T_\psi, t, \phi)$ for any finite distribution $\mathcal{T}$, under the usual two-time scale assumptions (Tang et al., 2023) and hence optimizing the **MC-UHM-JEPA** losses descends towards the best multi-linear factorization of the t-step transition matrices.

Similarly, as is shown in Bagatella et al. (2026), the density loss

$$\mathcal{L}_{\text{dens}}^{M^{\pi_z}}(\phi^M, T_z^M, \psi^M) = \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z}}||\phi^M T_z^M \psi^{M\top} - M^{\pi_z}||_F^2 \tag{30}$$

is a Lyapunov function of the **MC-JEPA** losses and hence optimizing the **MC-JEPA** losses descends towards the best multi-linear factorization of the successor measure matrix.

We consider the single policy case where $|\mathcal{Z}| = 1$, so that

$$\mathcal{L}_{\text{dens}}^{(P^{\pi_z})^t}(\phi, T_{z,t}, \psi) = \frac{1}{2}\mathbb{E}_{t\sim\mathcal{T}}||\phi T_{z,t}\psi^\top - (P^{\pi_z})^t||_F^2 \tag{31}$$

and

$$\mathcal{L}_{\text{dens}}^{M^{\pi_z}}(\phi^M, T_z^M, \psi^M) = \frac{1}{2}||\phi^M T_z^M \psi^{M\top} - M^{\pi_z}||_F^2 \tag{32}$$

for any $z \in \mathcal{Z}$. It is easy to see that that P^{π_z} and M^{π_z} share eigenvectors since $M = (I - \gamma P)^{-1}$ (Chandak et al., 2023). This also holds for any power of t. The Frobenius norms in Equation (29) and Equation (32) find the best d-rank approximations to $(P^{\pi_z})^t$ and M^{π_z} by selecting the d singular vectors corresponding to the d largest singular values. Under assumption A3, P^{π_z} (and hence M^{π_z}) are symmetric so their singular vectors are the eigenvectors and singular values are given by $|\lambda_z|^t$ and $|\frac{1}{1-\gamma\lambda_z}| = \frac{1}{1-\gamma\lambda_z}$, respectively, where λ_z are the eigenvalues of P^{π_z}. For any $\lambda \geq 0$, these orderings are consistent.

However, in the multi-policy case where $|\mathcal{Z}| > 1$, there is no guarantee without additional assumptions that P^{π_z} and $P^{\pi_{z'}}$ share an eigenbasis, so we cannot rely on a pure ranking argument. If we assume, however that P^{π_z} share a top-d subspace for all $z \in \mathcal{Z}$, then the optimal eigenvectors are shared across all $z \in \mathcal{Z}$ and the problem reduces to the single policy setting.

A.2 Proof of Theorem 2

We are interested in the fixed point of our latent-predictive loss Equation (16), which we rewrite here under the assumption of a finite MDP and linear predictors

$$\mathcal{L}_{\text{UHM-JEPA}}(\phi, T_{t,z}, \psi) = \begin{cases} \mathbb{E}_{(s,a,s')\sim\mathcal{D},z\sim\mathcal{Z}}||T_{\phi,t,z}^\top \phi(s) - \overline{\psi(s')}||_2^2 & \text{if } t = 1 \\ \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D},z\sim\mathcal{Z}, \\ a'\sim\pi_z(\cdot|s,a)}}||T_{\phi,t,z}^\top \phi(s) - \overline{T_\phi(\phi(s'), a', z, t-1)}||_2^2 & \text{if } t > 1 \end{cases} \tag{33}$$

Next, we write this in matrix form, using Assumption 1 ($D_\rho = I$):

$$\mathcal{L}_{\text{UHM-JEPA}}(\phi, T_{t,z}, \psi, t) = \begin{cases} \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z}}||\phi T_{t,z} - P^{\pi_z}\overline{\psi}||_F^2 & \text{if } t = 1 \\ \frac{1}{2}\mathbb{E}_{z\sim\mathcal{Z}}||\phi T_{t,z} - P^{\pi_z}\overline{\phi T_{t-1,z}}||_F^2 & \text{if } t > 1 \end{cases} \tag{34}$$

Taking the derivative with respect to $T_{t,z}$:

$$\nabla_{T_{t,z}} \mathcal{L}_{\text{UHM-JEPA}}(\phi, T_{t,z}, \psi) = \begin{cases} \phi^\top(\phi T_{t,z} - P^{\pi_z}\bar{\psi}) = T_{t,z} - \phi^\top P^{\pi_z}\psi & \text{if } t = 1 \\ \phi^\top(\phi T_{t,z} - P^{\pi_z}\overline{\phi T_{t-1,z}}) = T_{t,z} - \phi^\top P^{\pi_z}\overline{\phi T_{t-1,z}} & \text{if } t > 1 \end{cases} \tag{35}$$

where we use Assumption 2 ($\phi^\top \phi = I$)

Setting $\nabla_{T_{t,z}} \mathcal{L}_{\text{UHM-JEPA}}(\phi, T_{t,z}, \psi, t) = 0$ for all $t \geq 1$, we have:

$$T_{t,z} = \phi^\top P^{\pi_z}(\Pi_\phi P^{\pi_z})^{t-1}\psi \tag{36}$$

Clearly in the base case:

$$T^*_{1,z} = \phi^\top P^{\pi_z}\psi \tag{37}$$

Assume Equation (36) holds for $t-1$, so $T_{t-1,z} = \phi^\top P^{\pi_z}(\Pi_\phi P^{\pi_z})^{t-2}\psi$, and since we are at a stationary point, $T_{t-1,z} = \overline{T_{t-1,z}}$ then:

$$T_{t,z} = \phi^\top P^{\pi_z}\phi\overline{T_{t-1,z}} \tag{38}$$

$$= \phi^\top P^{\pi_z}\phi\phi^\top P^{\pi_z}(\Pi_\phi P^{\pi_z})^{t-2}\psi \tag{39}$$

$$= \phi^\top P^{\pi_z}(\Pi_\phi P^{\pi_z})^{t-1}\psi \tag{40}$$

$$\phi T_{t,z} = (\Pi_\phi P^{\pi_z})^t \psi \tag{41}$$

Therefore, by induction, Equation (36) holds for all $t \geq 1$.

A.3 Proof of Theorem 3

We show that Equation (18) is a contraction.

Consider the operator $\mathcal{T}^\nu$ corresponding to Equation (18):

$$(\mathcal{T}^\nu M)(s,a) = P(s,a) + \gamma \sum_{k\geq 1} \xi_\nu(k)\,(P^\pi)^{k+1}(s,a) + \gamma \sum_{k\geq 1} \nu(k)\,(P^\pi)^{k+1} M(s,a) \tag{42}$$

Let $\mathcal{M}$ be the space of bounded functions $M : S \times A \to \mathcal{M}(S)$ equipped with the norm

$$\|M\|_\infty := \sup_{(s,a)\in S\times A} \|M(s,a)\|_{TV}, \tag{43}$$

which is a Banach space. For any $M, M' \in \mathcal{M}$

$$\left\|(\mathcal{T}^\nu M(s,a) - \mathcal{T}^\nu M'(s,a)\right\|_{TV} \tag{44}$$

$$= \gamma \left\| \mathbb{E}\left[\sum_{k\geq 1} \nu(k)(M(s_{k+1},a_{k+1}) - M'(s_{k+1},a_{k+1}))|s_0 = s, a_0 = 0 \right] \right\|_{TV}$$

$$\leq \gamma \mathbb{E}\left[\sum_{k\geq 1} \nu(k) \left\|(M(s_{k+1},a_{k+1}) - M'(s_{k+1},a_{k+1}))\right\|_{TV} |s_0 = s, a_0 = a \right] \tag{45}$$

$$\leq \gamma \sum_{k\geq 1} \nu(k)\, \|(M - M')\|_\infty \tag{46}$$

$$\leq \gamma\, \|M - M'\|_\infty, \tag{47}$$

where the last line follows since ν is a sub-probability measure with $\sum_{k\geq 1} \nu(k) \leq 1$. Taking the supremum over (s,a) on the left-hand side gives

$$\|\mathcal{T}^\nu M - \mathcal{T}^\nu M'\|_\infty \leq \gamma\, \|M - M'\|_\infty. \tag{48}$$

Since $\gamma < 1$, $\mathcal{T}^\nu$ is a contraction, and by the Banach fixed-point theorem it admits a unique fixed point.

Next, we show that the fixed point of $\mathcal{T}^\nu$ is M^π

$$\mathcal{T}M^\pi = P + \gamma \sum_{k\geq 1} \xi_\nu(k)(P^\pi)^{k+1} + \gamma \sum_{k\geq 1} \nu(k)(P^\pi)^{k+1} M^\pi \tag{49}$$

$$= P + \gamma \sum_{k\geq 1} \left(\gamma^{k-1} - \sum_{i=1}^{k-1} \nu(i)\gamma^{k-i-1} \right) (P^\pi)^{k+1} + \gamma \sum_{k\geq 1} \nu(k)(P^\pi)^{k+1} M^\pi \tag{50}$$

$$= P + \sum_{k\geq 1} \gamma^k (P^\pi)^{k+1} - \gamma \sum_{k\geq 2} \sum_{i=1}^{k-1} \nu(i)\gamma^{k-i-1}(P^\pi)^{k+1} + \gamma \sum_{k\geq 1} \nu(k)(P^\pi)^{k+1} M^\pi \tag{51}$$

$$= \sum_{k\geq 0} \gamma^k (P^\pi)^{k+1} - \gamma \sum_{i\geq 1} \sum_{k\geq 1} \nu(i)\gamma^{k-1}(P^\pi)^{k+i+1} + \gamma \sum_{k\geq 1} \nu(k)(P^\pi)^{k+1} M^\pi \tag{52}$$

$$= M^\pi - \gamma \sum_{i\geq 1} \nu(i)(P^\pi)^{i+1} \sum_{k\geq 1} \gamma^{k-1}(P^\pi)^k + \gamma \sum_{k\geq 1} \nu(k)(P^\pi)^{k+1} M^\pi \tag{53}$$

$$= M^\pi - \gamma \sum_{i\geq 1} \nu(i)(P^\pi)^{i+1} M^\pi + \gamma \sum_{k\geq 1} \nu(k)(P^\pi)^{k+1} M^\pi \tag{54}$$

$$= M^\pi \tag{55}$$

A.4 Proof of Theorem 4

In the discrete case, we consider the action-free form of this equation

$$\mathcal{L}_{\text{MB-TD-JEPA}} = \frac{1}{2} \mathbb{E}_{\substack{s\sim\rho, z\sim\mathcal{Z}, s'\sim P^{\pi_z}(\cdot|s), \\ k\sim p_h, s^+\sim (P^{\pi_z})^k(\cdot|s')}} || T_{\phi,z}^\top \phi(s) - \overline{\psi(s')} - \gamma(w_k^\xi \overline{\psi(s^+)} + w_k^\nu \overline{T_{\phi,z}^\top \phi(s^+)}) ||_2^2 \tag{56}$$

We can write this in matrix form as

$$\mathcal{L}_{\text{MB-TD-JEPA}} = \frac{1}{2} \mathbb{E}_{z\sim\mathcal{Z}, k\sim p_h} || D^{1/2} \left(\phi T_{\phi,z} - P^{\pi_z} \overline{\psi} - \gamma (P^{\pi_z})^{k+1} (w_k^\xi \overline{\psi} + w_k^\nu \overline{\phi T_{\phi,z}}) \right) ||_F^2 \tag{57}$$

We take the derivative with respect to $T_{\phi,z}$:

$$\nabla_{T_{\phi,z}} \mathcal{L}_{\text{MB-TD-JEPA}} = \frac{1}{2} \mathbb{E}_{k\sim p_h} \left[\nabla_{T_{\phi,z}} || \left(\phi T_{\phi,z} - P^{\pi_z} \overline{\psi} - \gamma (P^{\pi_z})^{k+1} (w_k^\xi \overline{\psi} + w_k^\nu \overline{\phi T_{\phi,z}}) \right) ||_F^2 \right] \tag{58}$$

$$= \mathbb{E}_{k\sim p_h} \left[\phi^\top \left(\phi T_{\phi,z} - P^{\pi_z} \psi - \gamma (P^{\pi_z})^{k+1} (w_k^\xi \psi + w_k^\nu \phi T_{\phi,z}) \right) \right] \tag{59}$$

$$= \sum_{k\geq 1} p_h(k) \left(T_{\phi,z} - \phi^\top P^{\pi_z} \psi - \gamma \phi^\top (P^{\pi_z})^{k+1} (w_k^\xi \psi + w_k^\nu \phi T_{\phi,z}) \right) \tag{60}$$

$$= T_{\phi,z} - \phi^\top P^{\pi_z} \psi - \gamma \phi^\top \left(\sum_{k\geq 1} p_h(k) (P^{\pi_z})^{k+1} w_k^\xi \right) \psi - \tag{61}$$

$$\gamma \phi^\top \left(\sum_{k\geq 1} p_h(k) (P^{\pi_z})^{k+1} w_k^\nu \right) \phi T_{\phi,z} \tag{62}$$

$$= T_{\phi,z} - \phi^\top P^{\pi_z} \psi - \tag{63}$$

$$\gamma \phi^\top \left(\sum_{k\geq 1} \xi(k) (P^{\pi_z})^{k+1} \right) \psi - \gamma \phi^\top \left(\sum_{k\geq 1} \nu(k) (P^{\pi_z})^{k+1} \right) \phi T_{\phi,z} \tag{64}$$

$$\tag{65}$$

Setting $\nabla_{T_{\phi,z}} \mathcal{L}_{mb-td} = 0$

$$0 = T_{\phi,z} - \phi^\top P^{\pi_z} \psi - \gamma \phi^\top \left(\sum_{k \geq 1} \xi(k) (P^{\pi_z})^{k+1} \right) \psi - \gamma \phi^\top \left(\sum_{k \geq 1} \nu(k) (P^{\pi_z})^{k+1} \right) \phi T_{\phi,z} \tag{66}$$

$$\left(I - \gamma \phi^\top \left(\sum_{k \geq 1} \nu(k) (P^{\pi_z})^{k+1} \right) \phi) \right) T_{\phi,z} = \phi^\top \left(P^{\pi_z} + \gamma \left(\sum_{k \geq 1} \xi(k) (P^{\pi_z})^{k+1} \right) \right) \psi \tag{67}$$

$$T_{\phi,z} = \left(I - \gamma \phi^\top \left(\sum_{k \geq 1} \nu(k) (P^{\pi_z})^{k+1} \right) \phi) \right)^{-1} \phi^\top \left(P^{\pi_z} + \gamma \left(\sum_{k \geq 1} \xi(k) (P^{\pi_z})^{k+1} \right) \right) \psi \tag{68}$$

From Equation (18), we have:

$$M_\gamma^{\pi_z} = \left(I - \gamma \sum_{k \geq 1} \nu(k) (P^{\pi_z})^{k+1} \right)^{-1} \left(P^{\pi_z} + \gamma \sum_{k \geq 1} \xi(k) (P^{\pi_z})^{k+1} \right) \tag{69}$$

Then:

$$T_{\phi,z} = \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) M_\gamma^{\pi_z} \psi \tag{70}$$

where $A_\nu^{\pi_z} = \sum_{k \geq 1} \nu(k) (P^{\pi_z})^{k+1}$

Hence,

$$\phi T_{\phi,z} = \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) M_\gamma^{\pi_z} \psi \tag{71}$$

We can show that $\tilde{\Pi}_\phi = \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right)$ is idempotent:

$$\tilde{\Pi}_\phi^2 = \left(\phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) \right)^2 \tag{72}$$

$$= \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) \tag{73}$$

$$= \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \left(\phi^\top \phi - \gamma \phi^\top A_\nu^{\pi_z} \phi \right) \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) \tag{74}$$

$$= \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi \right) \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) \tag{75}$$

$$= \phi \left(I - \gamma \phi^\top A_\nu^{\pi_z} \phi) \right)^{-1} \phi^\top \left(I - \gamma A_\nu^{\pi_z} \right) \tag{76}$$

$$= \tilde{\Pi}_\phi \tag{77}$$

where we use Assumption A2 and hence $\phi^\top \phi = I$

Thus, we have that $\phi T_{\phi,z} = \tilde{\Pi}_\phi M_\gamma^{\pi_z} \psi$ where Π_ϕ is an oblique projection onto the span of ϕ

A.5 Connection to horizon consistency loss

We show that we can derive the horizon consistency loss from (Farebrother et al., 2026) from our ν-Bellman equation (Equation (18).

First, we multiply Equation (18) by $(1-\gamma)$ and let $\tilde{M}^\pi = (1-\gamma) M^\pi$ be the normalized successor measure.

$$\tilde{M}_\gamma^\pi = \tilde{M}_\gamma^\pi = (1-\gamma) P + \gamma (1-\gamma) \sum_{k \geq 1} \xi(k) (P^\pi)^{k+1} + \gamma \sum_{k \geq 1} \nu(k) (P^\pi)^{k+1} \tilde{M}_\gamma^\pi \tag{78}$$

Let $\nu(k)$ be the sub-probability measure $\nu(k) = (\gamma - \beta) \beta^{k-1}$

$$\xi_\nu(k) = \gamma^{k-1} - \sum_{i=1}^{k-1} \nu(i)\gamma^{k-i-1} \quad \forall \quad k \geq 1 \tag{79}$$

$$= \gamma^{k-1} - (\gamma-\beta)\gamma^{k-1}\beta^{-1}\sum_{i=1}^{k-1}(\frac{\beta}{\gamma})^i \tag{80}$$

$$= \gamma^{k-1} - (\gamma-\beta)\gamma^{k-2}\sum_{j=0}^{k-2}(\frac{\beta}{\gamma})^j \tag{81}$$

$$= \gamma^{k-1} - (\gamma-\beta)\frac{\gamma^{k-1}}{\gamma}\frac{1-(\frac{\beta}{\gamma})^{k-1}}{1-\frac{\beta}{\gamma}} \tag{82}$$

$$= \gamma^{k-1} - (\gamma-\beta)\frac{\gamma^{k-1}-\beta^{k-1}}{\gamma-\beta} \tag{83}$$

$$= \gamma^{k-1} - \gamma^{k-1} + \beta^{k-1} \tag{84}$$

$$= \beta^{k-1} \tag{85}$$

Plugging in to Equation (86)

$$\tilde{M}^\pi_\gamma = \tilde{M}^\pi_\gamma = (1-\gamma)P + \gamma(1-\gamma)\sum_{k\geq 1}\beta^{k-1}(P^\pi)^{k+1} + \gamma\sum_{k\geq 1}\nu(k)(P^\pi)^{k+1}\tilde{M}^\pi_\gamma \tag{86}$$

$$= \tilde{M}^\pi_\gamma = (1-\gamma)P + \gamma(1-\gamma)P^\pi\sum_{k\geq 1}\beta^{k-1}(P^\pi)^{k} + \gamma\sum_{k\geq 1}\nu(k)(P^\pi)^{k+1}\tilde{M}^\pi_\gamma \tag{87}$$

$$= \tilde{M}^\pi_\gamma = (1-\gamma)P + \gamma(1-\gamma)P^\pi M^\pi_\beta + \gamma\sum_{k\geq 1}\nu(k)(P^\pi)^{k+1}\tilde{M}^\pi_\gamma \tag{88}$$

$$= \tilde{M}^\pi_\gamma = (1-\gamma)P + \gamma\frac{1-\gamma}{1-\beta}P^\pi\tilde{M}^\pi_\beta + \gamma P^\pi\sum_{k\geq 1}\nu(k)(P^\pi)^{k}\tilde{M}^\pi_\gamma \tag{89}$$

$$\tag{90}$$

We rewrite this in terms of expectations:

$$\tilde{M}^\pi_\gamma(\cdot|s,a) = (1-\gamma)P(\cdot|s,a) + \gamma\frac{1-\gamma}{1-\beta}\mathbb{E}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}\left[\tilde{M}^\pi_\beta(\cdot|s',a')\right] \tag{91}$$

$$+\gamma\mathbb{E}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi(\cdot|s'),k\sim\nu(k),\\ s^+\sim(P^\pi)^k(\cdot|s',a'),\\ a^+\sim\pi(\cdot|s')}}\left[\tilde{M}^\pi_\gamma(\cdot|s^+,a^+)\right]$$

We want to estimate the final term as an expectation over samples from $\tilde{M}^\pi_\beta$ however $\tilde{M}^\pi_\beta$ implicitly samples $p(k) \sim (1-\beta)\beta^{k-1}$ while the expectation is over $\nu(k) = (\gamma-\beta)\beta^{k-1}$. Our importance weights are therefore $w_\nu = \frac{(\gamma-\beta)\beta^{k-1}}{(1-\beta)\beta^{k-1}} = \frac{\gamma-\beta}{1-\beta}$

Giving the final equation:

$$\tilde{M}^\pi_\gamma(\cdot|s,a) = (1-\gamma)P(\cdot|s,a) + \gamma\frac{1-\gamma}{1-\beta}\mathbb{E}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}\left[\tilde{M}^\pi_\beta(\cdot|s',a')\right] + \tag{92}$$

$$\gamma\frac{\gamma-\beta}{1-\beta}\mathbb{E}_{\substack{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s'),\\ s^+\sim\tilde{M}^\pi_\beta(\cdot|s',a'),\\ a^+\sim\pi(\cdot|s^+)}}\left[\tilde{M}^\pi_\gamma(\cdot|s^+,a^+)\right] \tag{93}$$

A.6 Derivation of importance weights for winsorized geometric distribution

Following Chung et al. (2026) we use a winsorized geometric distribution for training and sampling from our UHM (see Section B for more details on exact parameters). Here we show the derivation of the winsorized horizon weights which follows Chung et al. (2026) closely.

Let

$$\nu(i) := \begin{cases} (\gamma-\beta)\beta^{i-1}, & 1 \le i < k_{\max} \\ \gamma\beta^{k_{\max}-1}, & i = k_{\max} \\ 0 & i > k_{\max} \end{cases} \tag{94}$$

then

$$\begin{aligned} \xi_\nu(n) &= \gamma^{k-1} - (\gamma-\beta)\gamma^{k-1}\beta^{-1}\sum_{i=1}^{k-1}(\frac{\beta}{\gamma})^i \\ &= \gamma^{k-1} - (\gamma-\beta)\gamma^{k-2}\sum_{j=0}^{k-2}(\frac{\beta}{\gamma})^j \\ &= \gamma^{k-1} - (\gamma-\beta)\frac{\gamma^{k-1}}{\gamma}\frac{1-(\frac{\beta}{\gamma})^{k-1}}{1-\frac{\beta}{\gamma}} \\ &= \gamma^{k-1} - (\gamma-\beta)\frac{\gamma^{k-1}-\beta^{k-1}}{\gamma-\beta} \\ &= \gamma^{k-1} - \gamma^{k-1} + \beta^{k-1} \\ &= \beta^{k-1}, && 1 \le n \le k_{\max} \\ \xi_\nu(n) &= 0, && n > k_{\max} \end{aligned}$$

For proposal distribution p_h, we sample by taking $n' \sim \text{Geom}(1-\beta), \quad n = \min(n', k_{\max})$, giving importance weights

$$w^\xi(n) = \begin{cases} \frac{1}{1-\beta}, & n < k_{\max} \\ 1, & n = k_{\max} \end{cases} \tag{95}$$

$$w^\nu(n) = \begin{cases} \frac{\gamma-\beta}{1-\beta}, & n < k_{\max} \\ \gamma, & n = k_{\max} \end{cases} \tag{96}$$

B Experimental Details

B.1 Algorithm

In Algorithm 1, we provide the algorithm for asymmetric model-based TD-JEPA, where we color in blue components of the algorithm that differ from the standard TD-JEPA algorithm.

B.2 UHM Training Details

We provide a diagram of our latent-predictive UHM in Figure 3

Architecture All networks use a simple feedforward MLP structure. The UHM follows the same structure of the forward map in TD-JEPA which first feeds the task conditioning vector z, the state-action pair through separate MLP encoders before concatenating the outputs and feeding through a final MLP. The UHM augments the concatenated input to the final MLP with a sinusoidal time embedding.

Algorithm 1 Model-Based TD-JEPA for zero-shot RL

Inputs: Dataset $\mathcal{D}$, batch size B, regularization coefficient λ, UHM loss coefficient α, networks $\pi, T_\phi, T_\phi^{\text{UHM}}, \phi, T_\psi, T_\psi^{\text{UHM}}, \phi, \psi$
Initialize target networks: $T_\phi^- \leftarrow T_\phi,\ \phi^- \leftarrow \phi,\ T_\psi^- \leftarrow T_\psi,\ \psi^- \leftarrow \psi,\ T_\phi^{-\text{UHM}} \leftarrow T_\phi^{\text{UHM}}\ T_\psi^{-\text{UHM}} \leftarrow T_\psi^{\text{UHM}}$
while $n \leq N$ **do**
  ▷ *Sample training batch*
  $\{(s_i, a_i, s_i')\}_{i=1}^B \sim \mathcal{D}, \quad \{z_i\}_{i=1}^B \sim \mathcal{Z}, \quad \{a_i'\}_{i=1}^B \sim \{\pi(\phi^-(s_i'), z_i)\}_{i=1}^B$
  ▷ *Compute UHM losses*
  $\{t_i\}_{i=1}^B \sim p_H^{\text{train}}$
  $\{bs_i^{\phi,\psi}\}_{i=1}^B = \begin{cases} \psi^-(s_i') & \text{if } t_i = 1 \\ T_\phi^{-\text{UHM}}(\phi^-(s_i'), a_i', z_i, t_i - 1) & \text{if } t_i > 1 \end{cases}$
  $\{bs_i^{\psi,\phi}\}_{i=1}^B = \begin{cases} \phi^-(s_i') & \text{if } t_i = 1 \\ T_\psi^{-\text{UHM}}(\psi^-(s_i'), a_i', z_i, t_i - 1) & \text{if } t_i > 1 \end{cases}$
  $\widehat{\mathcal{L}}_{\text{UHM}}(\phi, T_\phi^{\text{UHM}}, \psi) = \frac{1}{2B} \sum_i \left\| T_\phi^{\text{UHM}}(\phi(s_i), a_i, z_i, t_i) - bs_i^{\phi,\psi} \right\|^2$
  $\widehat{\mathcal{L}}_{\text{UHM}}(\psi, T_\psi^{\text{UHM}}, \phi) = \frac{1}{2B} \sum_i \left\| T_\psi^{\text{UHM}}(\psi(s_i), a_i, z_i, t_i) - bs_i^{\psi,\phi} \right\|^2$
  ▷ *Compute ν-Bellman latent-predictive losses*
  $\{t_i\}_{i=1}^B \sim p_H(t)$
  $\tilde{\phi}(s_i^+) = T_\psi^{-\text{UHM}}(\psi^-(s_i'), a_i', z_i, t_i)$
  $\tilde{\psi}(s_i^+) = T_\phi^{-\text{UHM}}(\phi^-(s_i'), a_i', z_i, t_i)$
  $\{a_i^+\}_{i=1}^B \sim \{\pi(\tilde{\phi}(s_i^+), z_i\}_{i=1}^B$
  $\widehat{\mathcal{L}}_{\text{MB-TD-JEPA}}(\phi, T_\phi, \psi) = \frac{1}{2B} \sum_i \left\| T_\phi(\phi(s_i), a_i, z_i) - \psi^-(s_i') - \gamma(w_i^\xi \tilde{\psi}(s_i^+) + w_i^\nu T_\phi^-(\tilde{\phi}(s_i^+), a_i^+, z_i)) \right\|^2$
  $\widehat{\mathcal{L}}_{\text{MB-TD-JEPA}}(\psi, T_\psi, \phi) = \frac{1}{2B} \sum_i \left\| T_\psi(\psi(s_i), a_i, z_i) - \phi^-(s_i') - \gamma(w_i^\xi \tilde{\phi}(s_i^+) + w_i^\nu T_\psi^-(\tilde{\psi}(s_i^+), a_i^+, z_i)) \right\|^2$
  ▷ *Compute orthonormality regularization losses*
  $\widehat{\mathcal{L}}_{\text{REG}}(\phi) = \frac{1}{2B(B-1)} \sum_{i \neq j} \left(\phi(s_i)^\top \phi(s_j) \right)^2 - \frac{1}{B} \sum_i \phi(s_i)^\top \phi(s_i)$
  $\widehat{\mathcal{L}}_{\text{REG}}(\psi) = \frac{1}{2B(B-1)} \sum_{i \neq j} \left(\psi(s_i)^\top \psi(s_j) \right)^2 - \frac{1}{B} \sum_i \psi(s_i)^\top \psi(s_i)$
  ▷ *Compute actor loss*
  $\{\hat{a}_i\}_{i=1}^B \sim \{\pi(\phi(s_i), z_i)\}_{i=1}^B$
  $\widehat{\mathcal{L}}_{\text{actor}}(\pi) = -\frac{1}{B} \sum_{i=1}^B T_\phi(\phi(s_i), \hat{a}_i, z_i)^\top z_i$
  Update $\phi, T_\phi, T_\phi^{\text{UHM}}$ to minimize $\widehat{\mathcal{L}}_{\text{MB-TD-JEPA}}(\phi, T_\phi, \psi) + \lambda \widehat{\mathcal{L}}_{\text{REG}}(\phi) + \alpha \widehat{\mathcal{L}}_{\text{UHM}}(\phi, T_\phi^{\text{UHM}}, \psi)$
  Update $\psi, T_\psi, T_\psi^{\text{UHM}}$ to minimize $\widehat{\mathcal{L}}_{\text{MB-TD-JEPA}}(\psi, T_\psi, \phi) + \lambda \widehat{\mathcal{L}}_{\text{REG}}(\psi) + \alpha \widehat{\mathcal{L}}_{\text{UHM}}(\psi, T_\psi^{\text{UHM}}, \phi)$
  Update π to minimize $\widehat{\mathcal{L}}_{\text{actor}}(\pi)$
  Update target networks $\phi^-, T_\phi^-, \psi^-, T_\psi^-, T_\phi^{-\text{UHM}}, T_\psi^{-\text{UHM}}$ via EMA of $\phi, T_\phi, \psi, T_\psi, T_\phi^{\text{UHM}}, T_\psi^{\text{UHM}}$
end while

Training and sampling horizon distributions Following Chung et al. (2026), we use Winsorized Geometric Distributions for both training and sampling in the UHM. We provide a derivation of the importance weights for our choice of ν and p_H in Section A.6. We discuss the values of β and k_{max} in Section B.4.

Monte Carlo sampling with task dropout. To help ground the UHM in real data, we perform task dropout with Monte Carlo samples from the dataset for a portion of the sampled transitions in each training batch. Specifically, for a given transition (s_t, a_t, s_{t+1}) we sample a timestep $k \sim p_H^{\text{train}}$ and with probability ω, we take the state from the trajectory s_{t+k+1}, instead of computing the bootstrap target. For transitions in the batch where we use Monte Carlo sampling, we drop out the z conditioning in the UHM predictor, replacing with the batch average $\bar{z} = \sum_{i=1}^B z_i$.

B.3 Model-Based Updates Details

Blended model-based and standard targets In practice, we find it beneficial to blend model-based targets with standard TD-JEPA targets. Since both share the fixed point $M_\gamma^{\pi_z}$, blending the targets is valid. In particular, we use the blended targets

$$G_i^\phi = \psi^-(s_i') + \gamma \left(\zeta \left(w_i^\xi \tilde{\psi}(s_i^+) + w_i^\nu T_\phi^-(\tilde{\phi}(s_i^+), a_i^+, z_i) \right) + (1 - \zeta) T_\phi^-(\phi^-(s_i'), a_i', z_i') \right). \quad (97)$$

and

$$G_i^\psi = \phi^-(s_i') + \gamma \left(\zeta \left(w_i^\xi \tilde{\phi}(s_i^+) + w_i^\nu T_\psi^-(\tilde{\psi}(s_i^+), a_i^+, z_i) \right) + (1 - \zeta) T_\psi^-(\psi^-(s_i'), a_i', z_i') \right). \quad (98)$$

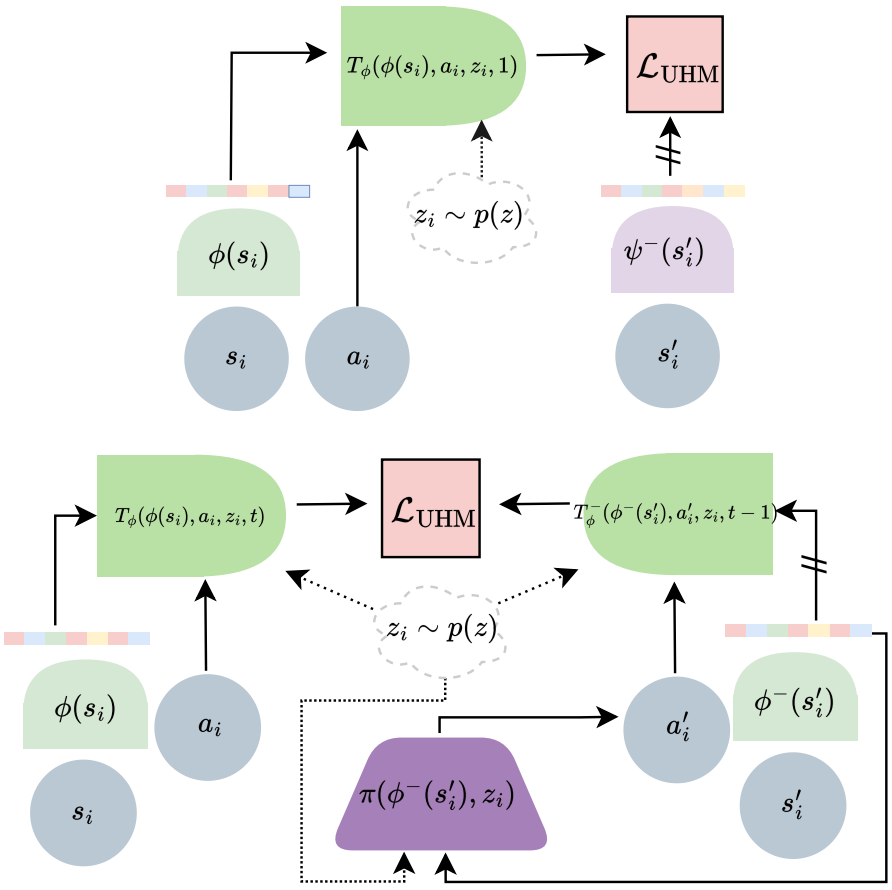


Figure 3: Diagram of the UHM training process. One step loss computation **(top)** trains the UHM predictor conditioned on representations ϕ to predict the next state representation in ψ. For $t > 1$ **(bottom)**, the UHM is trained for self consistency across the predictor conditioned on current state, action at time t and the predictor conditioned on next state, action and time $t-1$, Actions are drawn from the current policy π_z.

where ζ regulates the strength of model-based targets versus the standard TD-JEPA targets. We select the value of ζ through hyperparameter tuning as discussed in the following section.

Multi-sample estimates and ensembling To reduce variance in the model-based estimates, we find it beneficial to take multiple samples from the UHM and compute the bootstrap target for each of the samples. Further, we train ensemble over T_ϕ^{UHM}. For model-based updates, we take samples from all UHM in the ensemble and compute the bootstrap targets individually. The loss is then averaged over the TD error from the bootstraps computed with each individual UHM sample.

β-scheduling As was done in Chung et al. (2026), we use β-scheduling during training to increase the effective horizon of our sampling distribution for model-based updates as the UHM becomes more reliable. Specifically, we linearly ramp β over n_{warmup} steps, and select k_{max} as the 80th quantile of the resulting distribution. The values selected for k_{max} and n_{warmup} are discussed in the following section.

B.4 Hyperparameters and Tuning

MB-TD-JEPA-specific Hyperparameters In all experiments, the horizon distribution during training p_H^{train} is set to the winsorized geometric with $\beta = 0.95$ and $k_{\text{max}} = 30$. No β-scheduling is used

Table 3: Hyperparameters.

Hyperparameter	Value
Training	
Training steps	2×10^6
Batch size $\lvert\mathcal{B}\rvert$	256
Encoder target EMA rate τ_ϕ	0.005
Predictor target EMA rate τ_T	0.001
Model-based α warm-up steps	10^6
UHM samples per step	20
One-step auxiliary loss coefficient $\lambda_{\text{1-step}}$	1.0
Behavior cloning coefficient λ_{BC}	0.3
Training horizon distribution	
Geometric parameter β	0.95
Maximum horizon $k_{\max}$	30
UHM	
Learning rate η_{UHM}	10^{-4}
Target EMA rate τ_{UHM}	0.001
Architecture	
Representation dimension d	50
Normalize z	✓
Successor-measure predictor T_ϕ	
Hidden dimension	512
Hidden layers	2
Ensemble size	2
World-model predictor T_ϕ^{UHM}	
Hidden dimension	512
Hidden layers	2
Ensemble size	2
Time embedding	
Embedding dimension	128
Maximum period	512
Scale factor	16
Hidden dimension	256
Hidden layers	2
Actor π	
Hidden dimension	512
Hidden layers	2
Encoder ϕ	
Hidden dimension	512
Hidden layers	4
Layer normalization	✓
Decoder (detached)	
Hidden dimension	512
Hidden layers	2
Learning rate η_{dec}	10^{-4}

for the training distribution. We sweep over the sampling distribution p_H hyperparameters (n_{warmup} and β), as discussed below.

The task dropout ratio ω for training the UHM is set to 0.1 in all experiments. We sweep over the model-based vs standard mixing ratio ζ and the world model loss coefficient α, as discussed below.

TD-JEPA Hyperparameters All TD-JEPA hyperparameters, except those contained in the hyperparameter sweep in the following section, are kept consistent with Bagatella et al. (2026), except in the case of the predict target EMA update τ_T which is fixed at $\tau_T = 0.001$ for all experiments.

Hyperparameter Sweeps We sweep the hyperparameters for baseline TD-JEPA and TD-JEPA with auxiliary losses (same hyperparameter sweeps are used for one-step, UHM and UHM plus one-step)

as provided in Table 4. We add an additional orthonormal regularization parameter to the baseline sweep to make total number of swept configurations similar across methods.

Table 4: Hyperparameter sweeps for TD-JEPA baseline and auxiliary loss methods.

Method	$\lambda_{\phi,\text{ortho}}$	η_ϕ	α
TD-JEPA Baseline	[0.001, 0.01, 0.1]	[1e-4, 1e-5]	–
TD-JEPA + Auxiliary Loss	[0.001, 0.01]	[1e-4, 1e-5]	[0.001, 0.01]

In the model-based setting, at $\gamma = 0.995$ we select the best hyperparameters $(\lambda_{\phi,ortho}, \alpha, \eta_\phi)$ per environment from the UHM auxiliary loss sweep and then sweep over the model-based specific hyperparameters in Table 5.

Table 5: Hyperparameter sweeps for MB-TD-JEPA ($\gamma = 0.995$) with per-environment hyperparameters

				p_H		
Environment	$\lambda_{\phi,\text{ortho}}$	η_ϕ	α	β	n_{warmup}	ζ
`antmaze-medium-navigate`	0.01	10^{-4}	0.01	[0.9, 0.95]	[null, 10^6]	[0.2, 0.5]
`antmaze-medium-stitch`	0.001	10^{-5}	0.01	[0.9, 0.95]	[null, 10^6]	[0.2, 0.5]
`antmaze-medium-explore`	0.01	10^{-4}	0.01	[0.9, 0.95]	[null, 10^6]	[0.2, 0.5]
`antmaze-large-navigate`	0.01	10^{-4}	0.001	[0.9, 0.95]	[null, 10^6]	[0.2, 0.5]
`antmaze-large-stitch`	0.001	10^{-5}	0.01	[0.9, 0.95]	[null, 10^6]	[0.2, 0.5]
`antmaze-giant-navigate`	0.01	10^{-4}	0.01	[0.9, 0.95]	[null, 10^6]	[0.2, 0.5]

Finally, in the $\gamma = 0.999$ experiments, it was difficult to choose a best hyperparameter configuration from the auxiliary loss methods due to many configurations experiencing collapse which may be preventable with the model-based updates. For this reason, here we fix the p_H distribution to a β-schedule with $n_{\text{warmup}} = 10^6$ and $\beta = 0.95$ and sweep the hyperparameters in Table 6

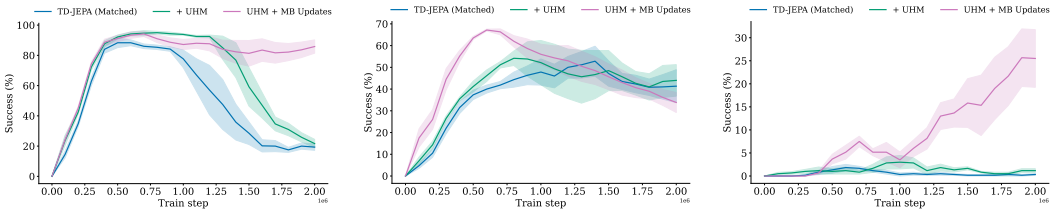
Table 6: Hyperparameter sweeps for MB-TD-JEPA ($\gamma = 0.999$), shared across all environments. $\lambda_{\phi,\text{ortho}}$ and η_ϕ are swept jointly as pairs.

Hyperparameter	**Values**
Orthonormality coefficient and ϕ learning rate $(\lambda_{\phi,\text{ortho}}, \eta_\phi)$	$[(0.001, 10^{-5}),\ (0.01, 10^{-4})]$
World-model loss coefficient α	$[0.001, 0.01]$
Horizon distribution p_H	
Winsorized geometric parameter β	0.95
Warm-up steps n_{warmup}	10^6
Model-based mixing ratio ζ	$[0.2, 0.5]$

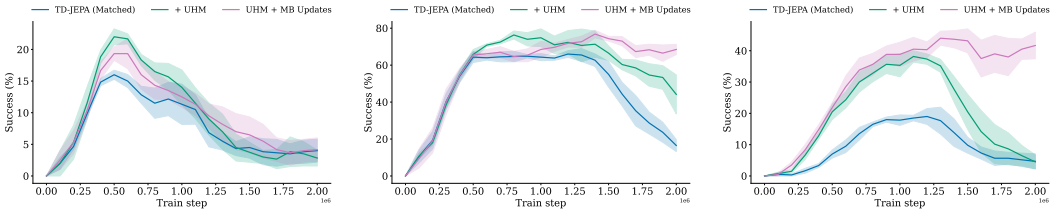
Across each method and environment, the best hyperparameter for reporting results is selected over a sweep of four seeds over these configurations. The best configuration is selected according to a lower confidence bound criteria, the mean less one standard error.

C Additional Results

We provide learning curves for success and TD loss for training at $\gamma = 0.999$ in all `antmaze` environments.

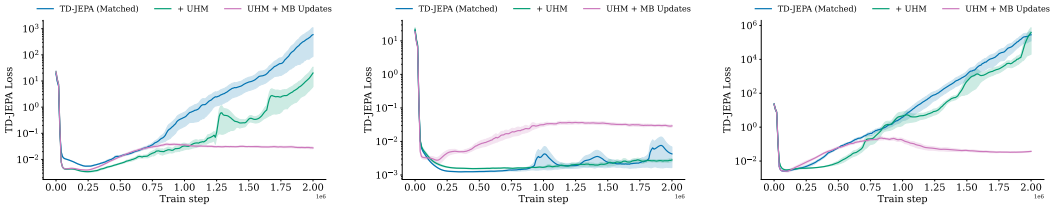


(a) Medium Navigate (b) Medium Stitch (c) Medium Explore

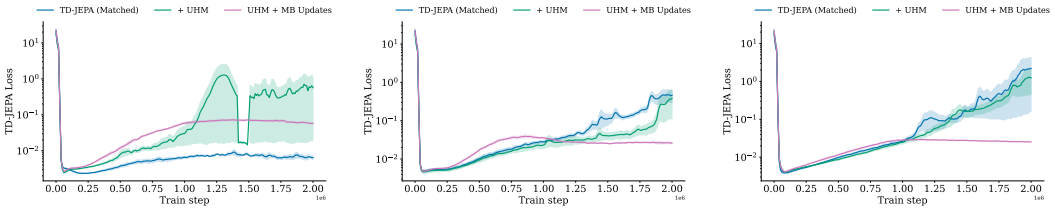


(d) Large Stitch (e) Large Navigate (f) Giant Navigate

Figure 4: Learning curves for average success across evaluation tasks for all `antmaze` environments. In these plots, for each method, we select the learning curve for the hyperparameters with the best ever performance during training instead of best final performance to better illustrate the catastrophic failure mode without model-based updates. Curves display a 3-step rolling average of averages over four seeds with standard error shading.



(a) Medium Navigate (b) Medium Stitch (c) Medium Explore



(d) Large Stitch (e) Large Navigate (f) Giant Navigate

Figure 5: Learning curves for TD-JEPA loss for all `antmaze` environments. In these plots, for each method, we select the learning curve for the hyperparameters with the best ever performance during training instead of best final performance to better illustrate the catastrophic failure mode without model-based updates. Curves display a 3-step rolling average of averages over four seeds with standard error shading.