MADE: A MEMORY-AUGMENTED DISTRIBUTION EXPANSION APPROACH VIA SLEEP CONSOLIDATION FOR CONTEXT-BASED OFFLINE META-REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Context-based offline meta-reinforcement learning acquires universal policies from fixed datasets by conditioning on task contexts, thereby enhancing agent generalization to unseen tasks. However, existing methods suffer from severe performance degradation under context shift, where test-time contexts are collected by behavior policies that differ from those seen during training. This brittleness arises because prior approaches optimize task representations solely on training data collected by fixed behavior policies, leaving agents vulnerable to out-of-distribution (OOD) failures at test time. Inspired by sleep consolidation in mammalian brains, we propose MADE, a memory-augmented framework that actively expands the task distribution in latent space during offline training, thereby enriching the diversity of context representations. MADE introduces three components: an uncertainty-aware memory bank identifying high-variance regions as expansion priorities, a counterfactual task generator synthesizing novel tasks via interpolation and conservative extrapolation, and a meta-critic filtering synthetic data through consistency, diversity, and causality assessments. Evaluations across five benchmarks show that MADE improves OOD performance by up to 13.10% compared to state-of-the-art methods and achieves up to 94.36% performance retention, effectively validating this active distribution expansion paradigm.

1 INTRODUCTION

Deep reinforcement learning (RL) (Kaelbling et al., 1996) enables autonomous agents to master complex tasks via environmental interaction, and has been widely adopted across both simulated and real-world scenarios. However, traditional RL often suffers from extreme sample inefficiency, typically requiring millions of trials to solve a single isolated task (Mnih et al., 2015; Hao et al., 2023; Wang et al., 2019). To overcome this, meta-reinforcement learning (meta-RL) distills transferable knowledge across related tasks to acquire strategies for rapid adaptation (Duan et al., 2016; Finn et al., 2017; Rakelly et al., 2019). By leveraging prior inductive biases to solve novel tasks with minimal experience (Finn et al., 2017; Beck et al., 2025), meta-RL facilitates diverse applications ranging from robotic manipulation (Schoettler et al., 2020) and personalized healthcare (Singh & Malhotra, 2023) to autonomous navigation under shifting dynamics (Wen et al., 2021).

However, meta-RL faces critical barriers in safety-critical scenarios where real-world exploration is prohibitively expensive (Luo et al., 2021; Levine et al., 2020), dangerous, or even infeasible, such as autonomous driving (Jaafra et al., 2019), financial trading (Urpí et al., 2021), and clinical healthcare (Smith et al., 2023). Offline meta-RL (OMRL) addresses this challenge by learning meta-policies exclusively from pre-collected datasets, eliminating the need for environment interaction during training (Hattori et al., 2023; Mitchell et al., 2021). Within the OMRL framework, context-based offline meta-RL (COMRL) has emerged as a dominant paradigm due to its computational efficiency and theoretical tractability (Rakelly et al., 2019). By mapping context trajectories to task embeddings that condition a universal policy, COMRL decouples task inference from execution to enable rapid test-time adaptation via latent representation updates instead of policy gradient steps.

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.

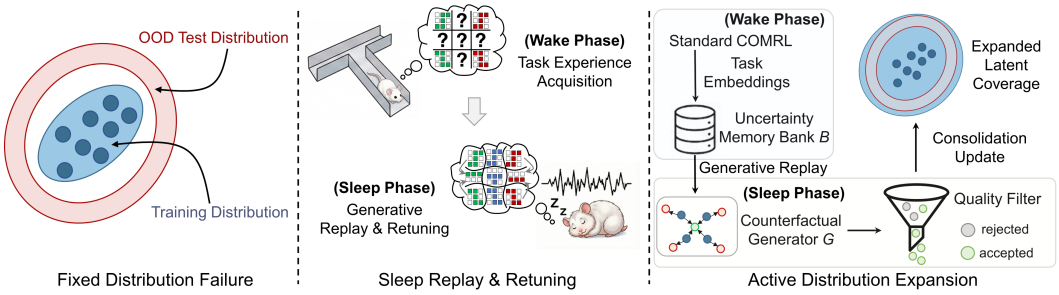


Figure 1: Conceptual illustration of Memory-Augmented Distribution Expansion. Left: Conventional context-based offline meta-learning fails to generalize when test tasks fall outside the fixed training distribution. Middle: Biological systems leverage sleep replay to actively remodel and retune internal representations beyond past experiences. Right: MADE implements active distribution expansion by consolidating uncertainty-aware counterfactual tasks into the latent space.

Current approaches such as FOCAL (Li et al., 2020), CORRO (Yuan & Lu, 2022), CSRO (Gao et al., 2023), and UNICORN (Li et al., 2024) optimize variants of the mutual information between task representations and identities, yet they remain limited by their reliance on fixed training distributions. While CERTAIN (Zhou et al., 2025) uses uncertainty awareness to reduce ambiguity, it requires environment interaction to correct inference errors. Static context trajectories in offline settings prevent such interaction, leading to severe performance degradation under out-of-distribution (OOD) shifts. Consequently, existing encoders struggle to generalize to unseen behavior policies, fundamentally restricting the utility of COMRL in real-world environments where distribution shift is inevitable.

Inspired by hippocampal sleep replay (Wilson & McNaughton, 1994; O'Neill et al., 2010; Klinzing et al., 2019), our framework leverages the spontaneous reactivation of neural patterns to facilitate active representation remodeling, thereby enhancing the agent's capacity to adapt to contexts generated by diverse behavior policies. Recent studies show that hippocampal activity during sleep predicts future place fields in novel environments, including representations that diverge from past experience (Maboudi et al., 2024). This biological mechanism suggests that adaptive systems can actively expand representational capacity during offline consolidation, rather than passively encoding a fixed distribution of experiences (Vaidya et al., 2025; Yaghoubi et al., 2026). Building on this principle, the dreams generated by MADE are rooted in the latent feature space acquired during training, simultaneously reflecting the structure of observed tasks and encoding predictive signals toward future task distributions, thus reducing overfitting to the fixed offline dataset while enhancing generalization under context shift.

In this paper, we propose Memory-Augmented Distribution Expansion (MADE), as illustrated in Figure 1. This framework overcomes fixed-distribution limitations by implementing computational analogs of sleep consolidation through three synergistic components. First, an uncertainty-aware memory bank tracks the distribution of task representations, identifying sparse regions in the latent space that indicate inadequate coverage of the task manifold. Second, a counterfactual generator synthesizes novel distributions by interpolating between existing prototypes to fill representational gaps, and extrapolating along uncertainty gradients to explore beyond observed task boundaries. Finally, a meta-critic evaluates these synthetic samples to ensure physical plausibility, filtering out low-fidelity distributions to ensure that only high-quality data contributes to the training process.

Our theoretical analysis relates test performance to latent distribution coverage under context shift, suggesting that MADE improves robustness by expanding task representations while filtering implausible synthetic contexts. Across five continuous control benchmarks, MADE consistently outperforms SOTA methods, achieving up to 94.36% performance retention under context-level shift with improvements of up to 13.10% in OOD scenarios. By shifting from passive encoding to active capacity expansion, this work defines a robust new direction for context-based offline meta-learning. Our main code has been anonymously uploaded to `https://anonymous.4open.science/r/made-314` without any author information.

2 Related Work

2.1 Context-Based Offline Meta-Reinforcement Learning

Context-based offline meta-reinforcement learning (COMRL) constitutes a predominant paradigm within the broader scope of offline meta-learning (OMRL). This framework decouples task inference from control to enable rapid adaptation without environment interaction. PEARL (Rakelly et al., 2019) pioneered the introduction of a probabilistic context encoder that maps trajectory segments into a latent task representation Z to improve performance. To enhance the discriminative power of these embeddings, subsequent research including FOCAL (Li et al., 2020), CORRO (Yuan & Lu, 2022), and CSRO (Gao et al., 2023) developed various contrastive and metric learning objectives. These methodological developments were formally unified by UNICORN (Li et al., 2024), which demonstrates that such objectives fundamentally operate as variational bounds for the mutual information $I(Z; M)$ between the representation Z and the task manifold M. While CERTAIN (Zhou et al., 2025) leverages uncertainty estimation to identify ambiguous contexts, it remains constrained by its inability to account for the underlying training distribution shift. Consequently, existing COMRL algorithms remain intrinsically limited by their reliance on the fixed training distribution, which restricts their capacity to generalize to tasks that lie beyond the support of the offline dataset.

2.2 Neuroscience-Inspired Consolidation Mechanisms

Biological systems exhibit remarkable meta-learning capabilities, facilitating rapid adaptation to novel environments through efficient knowledge transfer. This adaptability is partially underpinned by hippocampal memory consolidation, which facilitates the dynamic refinement of internal representations (Wilson & McNaughton, 1994; O'Neill et al., 2010; Buzsáki, 2015). However, emerging evidence indicates that hippocampal consolidation serves a more prospective function. Predictive retuning during sleep allows neural ensembles to pre-play trajectories that deviate from previous experience, effectively expanding the internal model to encompass potential future states (Maboudi et al., 2024; Sjøgård et al., 2025; Yaghoubi et al., 2026). While computational replay has been integrated into machine learning to mitigate catastrophic forgetting (Shin et al., 2017) or enhance sample efficiency (Hafner et al., 2019), these implementations primarily prioritize representational stability within the support of the training distribution. In contrast, MADE formalizes sleep consolidation as an active mechanism for distribution expansion. By translating biological predictive retuning into an information-theoretic framework, we enable agents to remodel their latent manifolds beyond the constraints of fixed offline datasets, thereby bridging the gap between biological flexibility and offline meta-learning.

3 Problem Formulation

Each task $M \sim \mathcal{P}(M)$ in reinforcement learning is modeled as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{T}_M, R_M, \rho_{0,M}, \gamma)$, where the transition dynamics, reward function, or initial-state distribution may vary across tasks. In context-based offline meta-reinforcement learning (COMRL), each task M is associated with a fixed offline dataset $\mathcal{D}_M$ collected by one or more behavior policies. A context trajectory $c \in \mathcal{C}$ sampled from $\mathcal{D}_M$ is encoded by a task encoder $q_\phi : \mathcal{C} \to \mathbb{R}^d$ into a latent representation $z = q_\phi(c)$, which conditions a universal policy $\pi_\theta(a \mid s, z)$.

The population performance objective can be written as

$$\max_{\phi,\theta} \mathbb{E}_{M\sim\mathcal{P}(M),\, c\sim p_{\mathcal{D}_M}(c),\, \tau\sim p_M^{\pi_\theta}(\tau|q_\phi(c))} \left[\sum_{t=0}^{H} \gamma^t r_t \right], \tag{1}$$

where H denotes the episode horizon, $p_{\mathcal{D}_M}(c)$ is the empirical context distribution induced by the offline dataset $\mathcal{D}_M$, and $p_M^{\pi_\theta}(\tau \mid q_\phi(c))$ is the trajectory distribution induced by task M and the conditioned policy $\pi_\theta(a \mid s, q_\phi(c))$. During training, both representation learning and policy optimization are performed using only the fixed offline datasets, without additional environment interaction.

A central goal of COMRL is to learn a representation $Z = q_\phi(C)$ that is informative about the underlying task identity while being robust to nuisance variation in the context. Recent methods

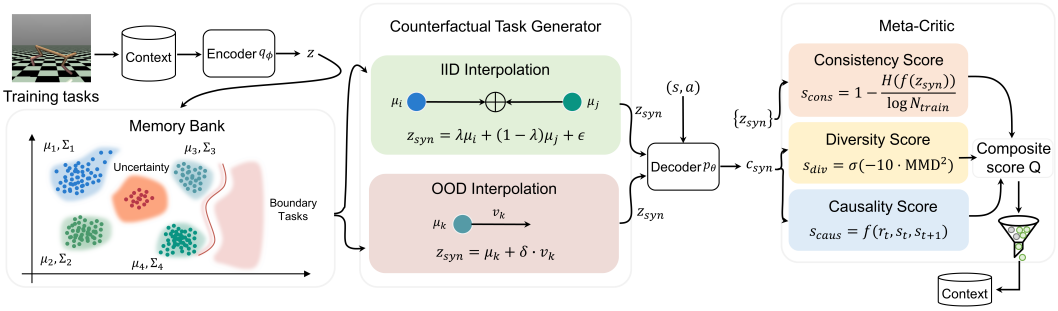


Figure 2: Overview of the MADE framework during the sleep phase. The Task Memory Bank (left) maintains latent distributions to identify sparse manifold regions. These guide the Counterfactual Task Generator (middle) in synthesizing IID and OOD latent tasks, which are decoded into synthetic contexts c_{syn}. Finally, the Meta-Critic (right) filters these samples based on a composite quality score Q, ensuring only high-fidelity data is used for memory consolidation.

optimize objectives related to the mutual information $I(Z; M)$ between task representations and task identities. UNICORN (Li et al., 2024) formalizes this view by decomposing each transition-level context variable into $X = (X_t, X_b)$, where $X_t = (s', r)$ contains task-relevant transition and reward information, while $X_b = (s, a)$ captures behavior-related state-action information. Under this decomposition, the task information captured by Z is bounded as

$$\begin{aligned} I(Z; X_t \mid X_b) \le I(Z; M) &\le I(Z; X_t \mid X_b) + I(Z; X_b) \\ &= I(Z; X). \end{aligned} \tag{2}$$

However, these objectives are typically optimized under the empirical training context distribution $p_{\text{train}}(c \mid M) = p_{\mathcal{D}_M}(c)$. At test time, the task parameter may remain within the training range while the context distribution changes, for example because contexts are generated by behavior policies different from those observed during training. We denote this shifted context distribution by $p_{\text{test}}(c \mid M)$. When $p_{\text{test}}(c \mid M) \neq p_{\text{train}}(c \mid M)$, the induced latent distribution $q_\phi(Z \mid M)$ may also shift, causing the conditioned policy $\pi_\theta(a \mid s, Z)$ to receive representations outside the region on which it was optimized. Thus, although existing information-theoretic objectives encourage task discriminability on the fixed offline dataset, they do not directly characterize how the learned representation behaves under shifted context distributions.

4 Method

4.1 Overview

MADE builds upon our framework by integrating an Uncertainty-Aware Memory Bank ($\mathcal{B}$) to identify under-represented latent regions, a Counterfactual Task Generator ($\mathcal{G}$) for interpolative and extrapolative expansion, and a Meta-Critic ($\mathcal{Q}$) for quality filtering (illustrated in Figure 2). The training procedure alternates between a *wake phase* for learning from fixed offline datasets and a *sleep phase* dedicated to expanding latent distribution coverage and consolidating the resulting synthetic contexts. The complete training algorithm and pseudocode are provided in Appendix A.

4.2 Uncertainty-Aware Memory Bank

The memory bank $\mathcal{B}$ encapsulates the distributional history of task representations to guide the expansion process. For each task M_i in the training set, $\mathcal{B}$ maintains a latent posterior $\mathcal{N}(\mu_i, \Sigma_i)$, which is updated recursively as the encoder q_ϕ encounters new context batches. Given a batch of contexts $\{c_j\}_{j=1}^{B}$ with embeddings $z_j = q_\phi(c_j)$, we compute the batch mean $\bar{z}_{\text{batch}}$ and sample

covariance S_{batch}, subsequently performing the following Bayesian updates:

$$\mu_i^{(t+1)} = (1-\kappa)\,\mu_i^{(t)} + \kappa\,\bar{z}_{\text{batch}}, \tag{3}$$

$$\Sigma_i^{(t+1)} = (1-\kappa)\,\Sigma_i^{(t)} + \kappa\,S_{\text{batch}} + \kappa(1-\kappa)\Big(\mu_i^{(t)} - \bar{z}_{\text{batch}}\Big)\Big(\mu_i^{(t)} - \bar{z}_{\text{batch}}\Big)^{\top}, \tag{4}$$

where $\kappa = B\,/\,(n_i^{(t)} + B)$ serves as a dynamic weighting factor proportional to the cumulative sample count $n_i^{(t)}$. This recursive formulation ensures that $\mathcal{B}$ captures the evolving representational manifold without the memory overhead of storing raw trajectories.

Task uncertainty is quantified by $u_i = \text{tr}(\Sigma_i)$, serving as a simple proxy for the local dispersion of task representations. High u_i values highlight unstable task identities, typically occurring near distribution boundaries or manifold gaps. $\mathcal{B}$ prioritizes these prototypes to expand latent coverage under potential context shift. Specifically, expansion is orchestrated via two mechanisms: (i) Stochastic Manifold Interpolation, which couples task pairs following $\mathcal{P}_{\text{pair}}(i,j) \propto \exp\big((u_i+u_j)/\tau\big)\cdot\exp\big(-\|\mu_i - \mu_j\|^2/2\ell^2\big)$, where τ controls uncertainty weighting and ℓ sets the Gaussian proximity bandwidth, filling local gaps without mixing distant tasks; and (ii) Frontier Extrapolation, which selects high-u_i tasks as directional seeds for conservative OOD exploration. This targeted expansion is designed to enlarge $p_{\text{aug}}(z \mid M)$ toward latent regions that may arise under shifted test contexts; direct distribution diagnostics appear in Appendix B.1.

4.3 Counterfactual Task Generator

The generator $\mathcal{G}$ operationalizes the augmented distribution $\mathcal{P}_{\text{syn}}$ by mapping latent expansions back to physically plausible transitions. This is achieved through a two-stage process: latent manifold expansion and trajectory synthesis. For Stochastic Manifold Interpolation, $\mathcal{G}$ targets representational gaps within the training support. Given a prototype pair (M_i, M_j) selected by $\mathcal{B}$, a synthetic latent is generated via a stochastic convex combination:

$$z_{\text{syn}} = \lambda\,\mu_i + (1-\lambda)\,\mu_j + \epsilon, \quad \lambda \sim \text{Beta}(\alpha_{\text{beta}}, \beta_{\text{beta}}), \tag{5}$$

where λ is the interpolation weight, α_{beta} and β_{beta} are Beta shape parameters, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is local diffusion noise. This mechanism ensures z_{syn} remains within the convex hull of the training data, effectively densifying the task manifold. To address OOD robustness, Frontier Extrapolation projects boundary prototypes M_k along their principal axes of uncertainty. We define the extrapolation direction v_k as the leading eigenvector of Σ_k and compute:

$$z_{\text{syn}} = \mu_k + \xi \cdot v_k, \quad \xi \sim \mathcal{TN}(\mu_\xi, \sigma_\xi^2;\, a,\, b), \tag{6}$$

where the step size ξ is drawn from a truncated normal distribution. This conservative shift allows z_{syn} to explore beyond the training support while avoiding distant latents that would loosen the transfer bound. Once the synthetic latents are sampled, $\mathcal{G}$ utilizes the decoder p_θ to synthesize corresponding context trajectories, effectively imagining the dynamics of counterfactual tasks. Starting from $s_0 \sim \rho_0$, a trajectory $c_{\text{syn}} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{n-1}$ is generated autoregressively:

$$(s_{t+1}, r_t) \sim p_\theta(\cdot \mid z_{\text{syn}}, s_t, a_t), \quad a_t \sim \mathcal{U}(\mathcal{A}), \tag{7}$$

where uniform action sampling is employed to maximize coverage of the synthetic transition space. This process yields N_{IID} and N_{OOD} synthetic contexts per sleep phase; Appendix B.2 evaluates their rollout fidelity.

4.4 Meta-Critic for Quality Assessment

Synthetic samples can fail in three distinct ways: semantic invalidity of z_{syn}, distributional misalignment of c_{syn}, and physically implausible dynamics. To safeguard the training manifold against low-fidelity expansions, the Meta-Critic $\mathcal{Q}$ implements a multi-dimensional filtering mechanism that maximizes the effective synthetic task count K_{syn}. Each sample $(z_{\text{syn}}, c_{\text{syn}})$ is evaluated via a composite score $\mathcal{Q} = w_{\text{cons}}\mathcal{S}_{\text{cons}} + w_{\text{div}}\mathcal{S}_{\text{div}} + w_{\text{caus}}\mathcal{S}_{\text{caus}}$, where the weights are normalized to unity. The Consistency Score $\mathcal{S}_{\text{cons}}$ validates the latent interpretability of z_{syn} using an auxiliary classifier f_ψ

trained on real embeddings: $\mathcal{S}_{\text{cons}} = 1 - H(f_\psi(z_{\text{syn}}))/\log N_{\text{train}}$. A low-entropy prediction ensures the synthetic latent resides in a coherent region of the task space. To promote distributional coverage, the Diversity Score $\mathcal{S}_{\text{div}}$ utilizes the Maximum Mean Discrepancy (MMD) (Smola et al., 2006) with an RBF kernel k to measure alignment with real contexts:

$$\text{MMD}^2 = \frac{1}{N_s^2}\sum_{i,j} k(c_i^s, c_j^s) + \frac{1}{N_r^2}\sum_{i,j} k(c_i^r, c_j^r) - \frac{2}{N_s N_r}\sum_{i,j} k(c_i^s, c_j^r). \tag{8}$$

where N_s and N_r denote the number of synthetic and real context samples respectively, $\{c_i^s\}_{i=1}^{N_s}$ denotes synthetic contexts decoded by p_θ, and $\{c_i^r\}_{i=1}^{N_r}$ denotes real context trajectories sampled from the offline dataset $\mathcal{D}$. The diversity metric $\mathcal{S}_{\text{div}} = \sigma(-10 \cdot \text{MMD}^2)$ ensures the generator explores the manifold without producing redundant or divergent samples. Complementing this, the Causality Score $\mathcal{S}_{\text{caus}}$ enforces physical plausibility via state-transition and reward-magnitude heuristics. We define the summation as:

$$\mathcal{S}_{\text{caus}} = \frac{1}{2n}\sum_{t=0}^{n-1}\Big[\exp(-\|s_{t+1} - s_t\|/\lambda_s) + \text{sigmoid}(\lambda_r - |r_t|)\Big], \tag{9}$$

where λ_s controls the sensitivity to state transition smoothness and λ_r defines the reward magnitude threshold. By penalizing abrupt state transitions and reward explosions, $\mathcal{S}_{\text{caus}}$ filters out hallucinated dynamics, ensuring that only synthetic tasks exceeding $\mathcal{Q} > \tau_{\text{thresh}}$ undergo Memory Consolidation; Appendix B.3 analyzes the score weights. The consolidation objective minimizes the reconstruction error of the synthetic task identity:

$$\mathcal{L}_{\text{cons}} = \|q_\phi(c_{\text{syn}}) - z_{\text{syn}}\|^2 + \beta\,\mathcal{L}_{\text{dec}}(z_{\text{syn}}, c_{\text{syn}}), \tag{10}$$

where $\mathcal{L}_{\text{dec}}$ is the supervised decoder loss and β is a scaling factor. This iterative refinement encourages the latent manifold to align with the augmented latent distribution $p_{\text{aug}}(z \mid M)$, thereby reducing training-test representation mismatch.

5 Experiments

We empirically demonstrate that MADE effectively mitigates performance degradation under distribution shift by actively expanding the task manifold. Our experiments are structured to address three primary objectives: (i) benchmarking performance against state-of-the-art COMRL methods on both in-distribution and out-of-distribution tasks; (ii) evaluating robustness across varying dataset qualities; and (iii) isolating the contributions of individual components via systematic ablations.

5.1 Experimental Setup

Environments. We evaluate MADE on five MuJoCo (Todorov et al., 2012) continuous control benchmarks: HalfCheetah-Dir, HalfCheetah-Vel, Ant-Dir, Hopper-Param, and Walker-Param. They require adaptation either across reward functions (HalfCheetah-Dir, HalfCheetah-Vel, Ant-Dir), or across dynamics via randomized physical properties (Hopper-Param, Walker-Param). The inherent context shift makes these environments standard benchmarks in the COMRL community, ensuring direct comparability with prior methods (Gao et al., 2023; Li et al., 2024; Zhou et al., 2025).

Data Collection. Following standard protocols in offline meta-RL, we generate datasets by independently training a Soft Actor-Critic agent (Haarnoja et al., 2018) for each task from random initialization. The complete replay buffer accumulated throughout the training process serves as the offline dataset for that task. This procedure naturally produces datasets with mixed behavior policies that span from exploratory random actions in early training to near-expert performance in later stages.

Baselines. We benchmark MADE against four state-of-the-art COMRL methods that represent the current frontier in addressing context shift and data augmentation: (i) CSRO (Gao et al., 2023), which utilizes contrastive metric learning to maximize task discriminability; (ii) ReDA (Jia et al., 2024), which employs adversarial augmentation to decouple task representations from behavior-policy bias; (iii) UNICORN (Li et al., 2024), a unified framework that optimizes mutual information bounds with reconstruction objectives; and (iv) CERTAIN (Zhou et al., 2025), which integrates uncertainty estimation for robust few-shot adaptation.

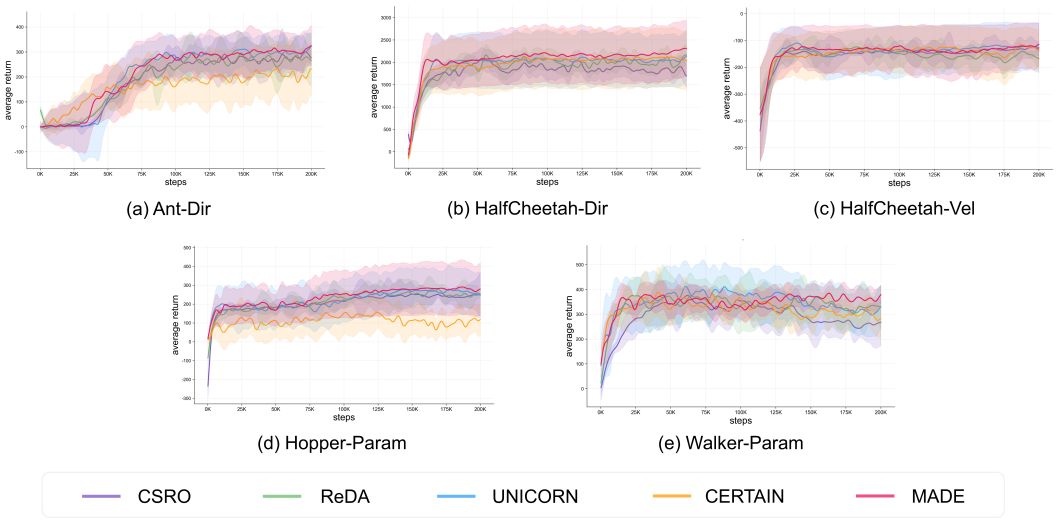


Figure 3: Testing returns of MADE against baselines on five benchmarks. Solid curves denote the mean performance across the task distribution, while shaded areas characterize the standard deviation of these trials.

Table 1: Performance comparison of MADE against state-of-the-art baselines under IID and OOD task distributions. Results are reported as mean ± standard deviation over four random seeds. The best performance is bolded and the second best is underlined.

ALGORITHM	ANT-DIR	HALFCHEETAH-DIR	HALFCHEETAH-VEL	HOPPER-PARAM	WALKER-PARAM
In-Distribution (IID) Results					
CSRO (GAO ET AL., 2023)	310.57±16.47	2175.33±47.74	-102.66±2.05	244.58±3.08	303.51±20.18
REDA (JIA ET AL., 2024)	317.31±12.29	2132.58±50.43	-107.53±5.22	253.09±5.26	329.55±24.43
UNICORN (LI ET AL., 2024)	330.35±11.12	2179.52±21.73	-86.37±2.03	242.95±3.15	352.71±23.77
CERTAIN (ZHOU ET AL., 2025)	245.69±17.26	2168.37±37.65	-106.42±3.72	173.66±15.23	315.22±15.80
MADE	**334.51±12.31**	**2318.14±33.62**	**-82.36±2.73**	**293.65±7.53**	**368.70±22.10**
Out-of-Distribution (OOD) Results					
CSRO (GAO ET AL., 2023)	273.18±11.29	1982.75±54.57	-131.40±3.53	235.41±5.37	278.58±22.25
REDA (JIA ET AL., 2024)	275.59±15.35	2085.36±51.81	-151.92±7.06	242.28±6.25	302.55±27.62
UNICORN (LI ET AL., 2024)	286.79±12.15	2001.44±24.81	-125.55±1.92	237.84±5.27	299.36±20.31
CERTAIN (ZHOU ET AL., 2025)	215.67±12.75	2053.21±39.12	-137.11±3.82	145.53±11.14	291.73±18.48
MADE	**293.53±11.17**	**2187.41±38.37**	**-118.01±1.42**	**274.02±4.15**	**322.15±22.73**

Implementation Details. To ensure a fair comparison, all methods share identical backbone architectures and hyperparameters where applicable. We report the mean performance and standard deviation across four independent random seeds to account for training stochasticity. Hyperparameters are provided in Appendix C; matched-budget controls and Meta-World results appear in Appendices B.4 and B.5.

5.2 MAIN RESULTS

5.2.1 COMPARISON WITH BASELINES

Following established evaluation protocols in COMRL (Gao et al., 2023; Li et al., 2024; Zhou et al., 2025), independent and identically distributed (IID) evaluation uses context trajectories from each test task's own replay buffer, where task parameters lie within the training range and behavior policies match the training agent. OOD evaluation addresses context-level shift where task parameters remain within the training range but contexts are collected by behavior policies from different tasks at various training stages, introducing a distribution mismatch at the context level.

Figure 3 illustrates that MADE achieves competitive or superior asymptotic performance across all benchmarks. In Hopper-Param (Figure 3d), MADE demonstrates a distinct advantage, reaching a

Table 2: Performance comparison under different data quality on Ant-Dir. Initial datasets contain trajectories from early training phases with exploratory behaviors, while Expert datasets consist of near-optimal demonstrations.

ALGORITHM	INITIAL		EXPERT	
	IID	OOD	IID	OOD
CSRO (GAO ET AL., 2023)	224.64±9.31	238.52±8.73	296.83±15.24	256.34±11.82
REDA (JIA ET AL., 2024)	307.13±13.54	277.21±14.63	319.42±12.61	279.73±15.32
UNICORN (LI ET AL., 2024)	318.27±10.62	281.35±11.28	330.82±11.14	291.03±11.71
CERTAIN (ZHOU ET AL., 2025)	234.92±13.71	214.43±11.23	242.81±15.93	218.94±12.52
MADE	**332.24±10.43**	**294.91±9.82**	**342.53±12.74**	**297.82±10.81**

Table 3: Test performance under extreme task-level OOD on Ant-Dir and Hopper-Param.

METHOD	ANT-DIR	HOPPER-PARAM
BASELINE (UNICORN)	249.73±18.54	194.52±15.37
RANDOM NOISE	241.74±16.82	183.76±9.82
UNICORN-DECODER	253.23±15.63	201.34±12.16
REDA-AUG	259.87±14.27	218.63±14.73
MADE	**272.68±11.93**	**247.81±5.42**

higher average return compared to the cluster of baselines. On directional tasks such as Ant-Dir and HalfCheetah-Dir, the algorithm maintains a consistent lead with narrower confidence intervals, which indicates that counterfactual task generation effectively densifies the latent manifold. While CERTAIN incorporates uncertainty estimation, its performance suggests that mechanisms primarily optimized for active online exploration may encounter limitations in strictly offline settings where environmental interaction is prohibited and the training distribution is fixed.

Table 1 quantifies the final testing performance and underscores the advantage of MADE under distribution shift. Under IID and OOD conditions, MADE achieves the highest returns in five benchmarks, notably exceeding the strongest baseline in Hopper-Param by 16.0%. In Hopper-Param, the OOD score of 274.02 represents a 13.10% improvement over the ReDA baseline. Similarly, MADE outperforms the strongest OOD baselines in Walker-Param and HalfCheetah-Dir by 6.5% and 4.9% respectively. Notably, MADE achieves the highest performance retention rate of 94.36% on HalfCheetah-Dir, demonstrating its ability to preserve performance under context-level shift. These consistent results across diverse benchmarks validate that active distribution expansion successfully reduces the performance gap between training and testing distributions, confirming that sleep consolidation enables generalization beyond fixed offline datasets. Extended experimental results are provided in Appendix B.

5.2.2 ROBUSTNESS TO DATA QUALITY

Table 2 evaluates robustness across varying dataset compositions. CSRO exhibits severe performance degradation under Initial distribution, as its contrastive learning mechanism struggles with the insufficient task discriminability of early exploration phases. ReDA and UNICORN demonstrate greater stability across data regimes, as their frameworks explicitly account for behavior policy diversity through adversarial augmentation and unified information objectives respectively. MADE consistently outperforms all baselines under both distributions, demonstrating robustness to data quality from exploratory to expert trajectories. The consistent superiority across data qualities validates that counterfactual task generation provides robust learning signals regardless of behavior policy characteristics, as synthetic tasks expand the manifold independent of trajectory optimality.

5.2.3 EXTREME TASK-LEVEL OOD EVALUATION

To validate the extrapolation capability of MADE under severe distribution shift, we conduct an extreme task-level OOD experiment on Ant-Dir and Hopper-Param following (Li et al., 2024), where training on consecutive middle-range tasks and testing on spatially separated tasks create a manifold gap with no neighboring training support. We compare five augmentation strategies:

Table 4: Ablation study on Ant-Dir and Walker-Param. Each variant removes one core component.

	Ant-Dir		Walker-Param	
Variant	IID	OOD	IID	OOD
MADE (Full)	**334.51±12.31**	**293.53±11.17**	**368.70±22.10**	**322.15±22.73**
w/o Memory Bank	319.41±16.54	274.18±15.23	317.84±26.31	291.47±24.82
w/o Interpolation	318.47±15.93	288.24±13.41	319.46±24.68	306.54±23.17
w/o Extrapolation	330.17±13.08	267.64±13.86	336.82±21.08	295.37±23.46
w/o Meta-Critic	303.94±18.82	271.23±17.63	297.83±28.42	283.21±27.15
w/o Consolidation	319.08±16.13	275.92±14.76	316.81±25.62	303.85±23.84
UNICORN (Baseline)	330.35±11.12	286.79±12.15	352.71±23.77	299.36±20.31

Baseline (no augmentation), Random Noise (Gaussian perturbations), UNICORN-Decoder (decoder-generated rollouts with perturbed latents $z + \epsilon$), ReDA-Aug (adversarial context augmentation), and MADE (uncertainty-guided extrapolation with meta-critic filtering). As shown in Table 3, MADE substantially outperforms all baselines in this extremely challenging scenario, achieving a 9.2% improvement over the UNICORN baseline on Ant-Dir, demonstrating that uncertainty-guided frontier extrapolation enables effective generalization beyond the convex hull of the training distribution. A consistent trend is observed on Hopper-Param, where MADE achieves a 27.4% gain over the UNICORN baseline, further corroborating that this advantage generalizes across environments with distinct task structures.

5.3 Ablation Study

Table 4 reveals the contribution of each component through systematic ablation. Walker-Param exhibits the same overall trend, with every component removal degrading performance and Meta-Critic removal causing the largest OOD drop, confirming consistent effects across task families. Meta-Critic removal causes the most severe degradation, with OOD performance dropping to 271.23, even below the UNICORN baseline, suggesting that unfiltered synthetic tasks may expand toward implausible latent regions. Removing the Memory Bank leads to significant OOD loss, demonstrating that uncertainty-guided expansion substantially outperforms random task generation. Interpolation and Extrapolation exhibit complementary roles: the former primarily maintains IID performance by densifying the manifold interior, while the latter explores distribution boundaries to improve OOD robustness. Notably, the dedicated consolidation mechanism (Eq. 10) proves critical on MADE, as its removal causes performance to fall below the UNICORN baseline, while the specialized sleep phase loss contributes an additional 6.4% OOD improvement through targeted latent space remodeling. Additional ablations, including hyperparameter sensitivity analysis and computational cost comparisons, are provided in Appendix B.

6 Conclusion and Limitations

This work addresses the brittleness of context-based offline meta-RL under distribution shift by introducing MADE, a framework that reframes meta-learning as active capacity expansion rather than passive encoding. Inspired by hippocampal sleep consolidation, MADE synthesizes high-quality counterfactual tasks through uncertainty-guided generation and quality filtering, thereby expanding latent task coverage and improving robustness under context distribution shift. Experiments across five benchmarks validate the effectiveness of this paradigm, demonstrating consistent improvements over state-of-the-art methods in both IID and OOD scenarios. By actively expanding the task manifold during offline training, MADE provides a robust approach for handling inevitable distribution shifts in real-world deployments. Despite its effectiveness, MADE has several limitations. First, the theoretical analysis offers motivational rather than complete support. Second, synthetic quality depends on decoder fidelity, which may degrade under extremely sparse data even with meta-critic filtering. Third, following standard COMRL protocols for fair comparison, validation is confined to MuJoCo locomotion, leaving hierarchical, manipulation, and multi-agent settings unexplored. Future work will pursue online adaptation, broader domain validation, and integration with world models.

AI USE STATEMENT

Large language models (LLMs) were used solely for language polishing to improve the clarity and readability of this manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A tutorial on meta-reinforcement learning. *Foundations and Trends in Machine Learning*, 18(2-3):224–384, 2025.

György Buzsáki. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, 25(10):1073–1188, 2015.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, et al. Context shift reduction for offline meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 36:80024–80043, 2023.

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

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Jianye Hao, Tianpei Yang, Hongyao Tang, et al. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):8762–8782, 2023.

Ryoma Hattori, Natalie G. Hedrick, Aditi Jain, et al. Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, 26(12):2182–2191, 2023.

Yesmina Jaafra, Aline Deruyver, Jean Luc Laurent, and Mohamed Saber Naceur. Context-aware autonomous driving using meta-reinforcement learning. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 450–455. IEEE, 2019.

Chengxing Jia, Fuxiang Zhang, Yi-Chen Li, Chen-Xiao Gao, Xu-Hui Liu, Lei Yuan, Zongzhang Zhang, and Yang Yu. Disentangling policy from offline task representation learning via adversarial data augmentation. *arXiv preprint arXiv:2403.07261*, 2024.

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

Jens G. Klinzing, Niels Niethard, and Jan Born. Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, 22(10):1598–1610, 2019.

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

Lanqing Li, Rui Yang, and Dijun Luo. Focal: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. *arXiv preprint arXiv:2010.01112*, 2020.

Lanqing Li, Hai Zhang, Xinyu Zhang, Shatong Zhu, Yang Yu, Junqiao Zhao, and Pheng-Ann Heng. Towards an information theoretic framework of context-based offline meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 37:75642–75667, 2024.

Michael Luo, Ashwin Balakrishna, Brijen Thananjeyan, Suraj Nair, Julian Ibarz, Jie Tan, Chelsea Finn, Ion Stoica, and Ken Goldberg. Mesa: Offline meta-rl for safe adaptation and fault tolerance. *arXiv preprint arXiv:2112.03575*, 2021.

Kourosh Maboudi, Bapun Giri, Hiroyuki Miyawaki, Caleb Kemere, and Kamran Diba. Retuning of hippocampal representations during sleep. *Nature*, 629(8012):630–638, 2024.

Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline meta-reinforcement learning with advantage weighting. In *International Conference on Machine Learning*, pp. 7780–7791. PMLR, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Joseph O'Neill, Barty Pleydell-Bouverie, David Dupret, and Jozsef Csicsvari. Play it again: reactivation of waking experience and memory. *Trends in neurosciences*, 33(5):220–229, 2010.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pp. 5331–5340. PMLR, 2019.

Gerrit Schoettler, Ashvin Nair, Juan Aparicio Ojea, Sergey Levine, and Eugen Solowjow. Meta-reinforcement learning for robotic industrial insertion tasks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9728–9735. IEEE, 2020.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Kuljeet Singh and Deepti Malhotra. Meta-health: Learning-to-learn (meta-learning) as a next generation of deep learning exploring healthcare challenges and solutions for rare disorders: A systematic analysis. *Archives of Computational Methods in Engineering*, 30(7), 2023.

M. Sjøgård, B. Baxter, D. Mylonas, et al. Hippocampal ripples predict motor learning during brief rest breaks in humans. *Nature Communications*, 16(1):6089, 2025.

Benjamin Smith, Anahita Khojandi, and Rama Vasudevan. Bias in reinforcement learning: A review in healthcare applications. *ACM Computing Surveys*, 56(2):1–17, 2023.

Alex J. Smola, Arthur Gretton, and Karsten M. Borgwardt. Maximum mean discrepancy. In *13th International Conference on Neural Information Processing*, 2006.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.

S. P. Vaidya, G. Li, R. A. Chitwood, et al. Formation of an expanding memory representation in the hippocampus. *Nature Neuroscience*, 28(7):1510–1518, 2025.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, et al. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.

Shuhuan Wen, Zeteng Wen, Di Zhang, Hong Zhang, and Tao Wang. A multi-robot path-planning algorithm for autonomous navigation using meta-reinforcement learning based on transfer learning. *Applied Soft Computing*, 110:107605, 2021.

Matthew A Wilson and Bruce L McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994.

Mohammad Yaghoubi, M Ganesh Kumar, Andres Nieto-Posadas, Coralie-Anne Mosser, Thomas Gisiger, Émmanuel Wilson, Cengiz Pehlevan, Sylvain Williams, and Mark P Brandon. Predictive coding of reward in the hippocampus. *Nature*, pp. 1–7, 2026.

Haoqi Yuan and Zongqing Lu. Robust task representations for offline meta-reinforcement learning via contrastive learning. In *International Conference on Machine Learning*, pp. 25747–25759. PMLR, 2022.

Hongtu Zhou, Ruiling Yang, Yakun Zhu, Haoqi Zhao, Hai Zhang, Di Zhang, Junqiao Zhao, Chen Ye, and Changjun Jiang. CERTAIN: Context uncertainty-aware one-shot adaptation for context-based offline meta reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=cx9AnB8rzi.

A Training Algorithm and Sleep Consolidation

The complete MADE training procedure alternates between wake and sleep phases to bridge the gap between static training data and OOD test environments. During wake phases, the agent performs standard offline meta-RL updates: contexts are mapped to task representations via q_ϕ, and the universal policy π_θ is optimized through established objectives. Concurrently, the Uncertainty-Aware Memory Bank tracks representational variance using recursive Bayesian estimation (Eqs. 7–8).

Every T_{replay} iterations, the system enters the sleep phase to operationalize distribution expansion. The Counterfactual Task Generator synthesizes N_{IID} and N_{OOD} tasks via latent interpolation and extrapolation, which are then filtered by the Meta-Critic to ensure high fidelity and causality. These high-quality samples $(z_{\text{syn}}, c_{\text{syn}})$ are processed during Memory Consolidation, minimizing the following objective:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}_{\mathcal{D}_{\text{syn}}}\left[\|q_\phi(c_{\text{syn}}) - z_{\text{syn}}\|^2 + \beta\, \mathcal{L}_{\text{dec}}(z_{\text{syn}}, c_{\text{syn}})\right]. \tag{11}$$

Algorithm 1 MADE Training with Sleep Consolidation

Input: Offline datasets $\{\mathcal{D}_M\}$, encoder q_ϕ, policy/decoder π_θ / p_θ, meta-critic f_ψ
Initialize memory bank $\mathcal{B} \leftarrow \emptyset$
for $t = 1$ **to** T_{total} **do**
  // Wake Phase: Standard COMRL Training
  Sample tasks $\{M_i\} \sim \mathcal{P}_{\text{train}}$ and contexts $\{c_i \sim \mathcal{D}_{M_i}\}$
  Encode $z_i \leftarrow q_\phi(c_i)$ and update $\mathcal{B}$ via recursive Bayesian update (Eqs. 8–9)
  Update ϕ, θ via offline meta-RL objectives
  if $t \bmod T_{\text{replay}} = 0$ **then**
    // Sleep Phase: Distribution Expansion
    $\mathcal{Z}_{\text{syn}} \leftarrow \mathcal{G}.\text{generate}(\mathcal{B}, N_{\text{IID}}, N_{\text{OOD}})$ (Eqs. 10–11)
    $\mathcal{D}_{\text{syn}} \leftarrow \{(z, c) : z \in \mathcal{Z}_{\text{syn}},\ c \sim p_\theta(\cdot|z, \cdot, \cdot),\ \mathcal{Q}(z, c) > \tau_{\text{thresh}}\}$
    for $k = 1$ **to** K_{cons} **do**
      Sample $(z_{\text{syn}}, c_{\text{syn}}) \sim \mathcal{D}_{\text{syn}}$
      Update ϕ, θ via $\nabla_{(\phi,\theta)}\mathcal{L}_{\text{cons}}$ (Eq. 15)
    end for
  end if
end for
Return: Encoder q_ϕ, policy π_θ

B Additional Experimental Results

B.1 Distribution-Level Validation

We directly evaluate whether MADE moves the augmented latent distribution toward independently held-out test contexts. The test contexts were collected by behavior policies excluded from training and memory-bank construction. Table 5 reports the Wasserstein-1 distance and squared MMD from p_{train} and p_{aug} to p_{test} over three random seeds.

Table 5: Latent distribution distance to independently held-out test contexts. Results are mean ± standard deviation over three seeds.

Environment	W_1(train,test)	W_1(aug,test)	Reduction	MMD^2(train,test)	MMD^2(aug,test)	Reduction
Ant-Dir	0.2677±0.0079	0.2597±0.0170	2.99%	0.0227±0.0031	0.0162±0.0036	28.5%
HalfCheetah-Dir	0.3310±0.0035	0.2376±0.0065	28.2%	0.0462±0.0070	0.0233±0.0020	49.7%
HalfCheetah-Vel	0.6807±0.0642	0.5484±0.0441	19.4%	0.0361±0.0037	0.0216±0.0079	40.1%
Hopper-Param	2.1060±0.0581	1.8405±0.0572	12.6%	0.0354±0.0031	0.0247±0.0069	30.2%
Walker-Param	2.1125±0.0522	1.8591±0.0705	12.0%	0.0826±0.0057	0.0521±0.0026	36.8%

Both discrepancies decrease after augmentation in every environment. The W_1 reduction ranges from 2.99% to 28.2%, while MMD^2 falls by 28.5%–49.7%. Ant-Dir has a modest W_1 change but a clear MMD^2 reduction, whereas the other environments improve substantially under both metrics. These results empirically validate the distributional effect targeted by MADE without treating it as a theoretical guarantee.

Table 6: Post-hoc rollout validation of synthetic contexts. Accepted values show the range across IID and OOD synthetic samples.

Environment	Real baseline	Accepted (IID/OOD)	Rejected ($\mathcal{Q} < \tau$)	Spearman r ($\mathcal{Q}$ vs. divergence)
Ant-Dir	7.0±0.6	7.3–8.1	11.2±1.4	-0.42 ($p < 0.01$)
HalfCheetah-Dir	4.0±0.8	4.5–5.0	7.8±1.3	-0.48 ($p < 0.01$)
HalfCheetah-Vel	3.9±0.7	4.3–4.9	7.2±1.5	-0.38 ($p < 0.01$)
Hopper-Param	1.0±0.5	1.1–1.3	2.1±0.9	-0.45 ($p < 0.01$)
Walker-Param	18.0±4.0	19.8–21.5	29.7±5.9	-0.40 ($p < 0.01$)
Pooled (Fisher z)	–	–	–	-0.43 ($p < 0.001$)

B.2 Synthetic Context Rollout Validity

We assess decoder-generated contexts against real-environment rollouts. For each synthetic latent, its source task parameters are reconstructed and evaluated with a fixed open-loop action sequence. Table 6 compares multi-step rollout divergence for real held-out reconstruction, accepted synthetic samples, and samples rejected by the Meta-Critic.

Accepted samples remain close to the real reconstruction baseline, whereas rejected samples have consistently larger divergence. The quality score correlates negatively with rollout error in all five environments, with a pooled correlation of -0.43. Thus, the composite Meta-Critic score tracks external dynamics error and removes many high-divergence samples.

Table 7: Sensitivity to Meta-Critic score weights on Ant-Dir.

w_{cons}	w_{div}	w_{caus}	IID Return	OOD Return
1.00	0.00	0.00	320.53±14.28	281.69±13.92
0.00	1.00	0.00	316.85±15.61	285.25±14.10
0.00	0.00	1.00	313.26±16.84	288.41±13.24
0.50	0.25	0.25	328.98±13.13	289.74±12.46
0.25	0.50	0.25	325.62±14.32	291.39±11.97
0.25	0.25	0.50	323.81±14.70	292.17±11.62
0.33	0.33	0.33	**334.51±12.31**	**293.53±11.17**

B.3 Meta-Critic Weight Sensitivity

The balanced default achieves the strongest IID and OOD returns, while every single-score or skewed configuration performs worse (Table 7). Among single-score variants, causality-only filtering gives the highest OOD return, showing that $\mathcal{S}_{\text{caus}}$ carries a useful signal; its gap to the balanced model shows that consistency and diversity provide complementary information.

B.4 Matched-Budget Augmentation Controls

To distinguish memory-guided expansion from generic augmentation, all controls generate the same number of synthetic samples per sleep phase ($N_s = 64$). Gaussian Noise and Latent Mixup remove uncertainty-guided targeting; w/o Memory Bank changes only the task-pairing policy from uncertainty-prioritized to random pairing; UNICORN-Decoder tests decoder augmentation without MADE's full expansion pipeline (Table 8).

Gaussian Noise, Latent Mixup, and random pairing all underperform the no-augmentation UNICORN baseline across the five environments. UNICORN-Decoder provides small gains but remains below full MADE, especially on HalfCheetah-Dir, Hopper-Param, and Walker-Param. Because every control uses the same synthetic sample budget, these results rule out data volume or decoder use alone and isolate the benefit of memory-guided targeting within the complete pipeline.

Table 8: Matched-budget augmentation controls across five environments. Results are mean ± standard deviation.

(a) IID returns

Method	Ant-Dir	HC-Dir	HC-Vel	Hopper-P	Walker-P
Gaussian Noise	301.79±14.50	2041.55±42.30	−93.56 ± 4.20	222.66±9.10	331.05±25.60
Latent Mixup	311.02±13.80	2074.11±40.10	−91.16 ± 3.95	229.83±8.60	338.14±24.30
w/o Memory Bank	319.41±16.54	2111.34±41.50	−90.12 ± 3.80	233.36±8.40	317.84±26.31
UNICORN (no aug.)	330.35±11.12	2179.52±21.73	−86.37 ± 2.03	242.95±3.15	352.71±23.77
UNICORN-Decoder	331.51±12.80	2209.60±38.50	−85.58 ± 3.20	248.54±7.80	356.77±22.90
MADE (Full)	**334.51±12.31**	**2318.14±33.62**	**−82.36 ± 2.73**	**293.65±7.53**	**368.70±22.10**

(b) OOD returns

Method	Ant-Dir	HC-Dir	HC-Vel	Hopper-P	Walker-P
Gaussian Noise	262.00±15.34	1875.00±48.26	−136.00 ± 5.18	218.00±10.42	281.00±27.15
Latent Mixup	270.00±14.27	1905.00±45.63	−132.50 ± 4.61	225.00±9.38	287.00±25.04
w/o Memory Bank	274.18±15.23	1938.60±43.50	−131.00 ± 4.85	228.50±9.90	291.47±24.82
UNICORN (no aug.)	286.79±12.15	2001.44±24.81	−125.55 ± 1.92	237.84±5.27	299.36±20.31
UNICORN-Decoder	287.80±13.00	2029.00±40.00	−124.40 ± 3.80	243.30±8.00	302.80±23.00
MADE (Full)	**293.53±11.17**	**2187.41±38.37**	**−118.01 ± 1.42**	**274.02±4.15**	**322.15±22.73**

Table 9: Performance on Meta-World Reach over four random seeds.

Algorithm	IID Return	OOD Return
CSRO	2997.42±235.67	2685.31±182.44
ReDA	3083.19±248.52	2803.76±196.28
UNICORN	3150.84±211.93	2780.55±178.61
CERTAIN	2636.27±263.15	2338.09±214.37
MADE	**3410.63±229.80**	**3020.94±171.52**

B.5 Cross-Domain Generalization on Meta-World Reach

We evaluate MADE beyond MuJoCo locomotion on Meta-World Reach. The task controls a 7-DoF Sawyer arm to reach continuously sampled 3D goals, changing both the embodiment and task geometry and using a 500-step context horizon. We retain the main context-shift protocol: IID contexts come from the same task's behavior policy, while OOD contexts come from other tasks and earlier policy stages.

As shown in Table 9, MADE achieves the best IID and OOD returns, improving OOD performance by 7.7% over the strongest baseline, ReDA. This result extends the main trend to manipulation with spatial goal variation, providing evidence that MADE's gains can generalize across environments.

B.6 Training Dynamics Analysis

Figure 4 presents the training curves across all five benchmarks. MADE demonstrates stable convergence and achieves competitive performance in most environments, particularly excelling in Hopper-Param where distribution expansion provides substantial gains. In Walker-Param, UNICORN attains higher training returns, likely due to the large parameter variance in this environment requiring more refined base representations that benefit from UNICORN's unified objective. Overall, the training dynamics confirm that the distribution expansion mechanism of MADE provides effective learning signals while maintaining stability comparable to baseline methods.

B.7 Hyperparameter Sensitivity Analysis

As shown in Table 10, the sleep phase frequency exhibits a trade-off between training stability and expansion coverage on Ant-Dir. Infrequent replay misses opportunities for timely manifold refinement, while the optimal frequency at $T_{\text{replay}} = 50$ balances stability and coverage. Performance saturates around 64 synthetic tasks per phase, with diminishing returns beyond 80 due to redundancy in the expanded distribution.

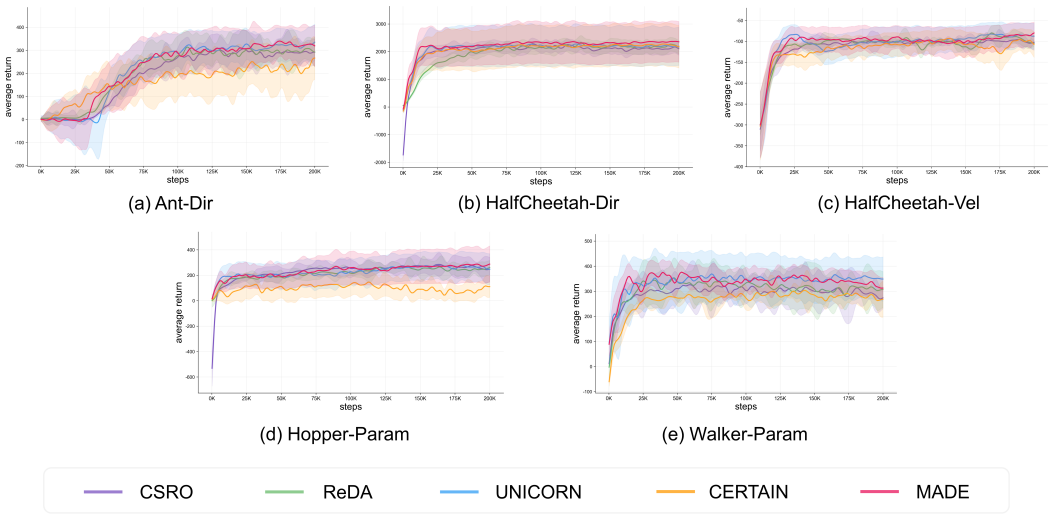


Figure 4: Training returns of MADE against baselines on five benchmarks. Solid curves denote the mean performance across the task distribution, while shaded areas characterize the standard deviation of these trials.

Table 10: Ablation on sleep phase configuration on Ant-Dir.

T_{REPLAY}	TOTAL SAMPLES	IID RETURN	OOD RETURN
10	64	325.83±12.74	281.46±10.53
25	64	330.47±11.86	288.37±11.62
50	64	**334.51±12.31**	**293.53±11.17**
100	64	332.28±12.23	290.18±9.34
200	64	327.94±13.65	284.73±10.82
50	32	328.63±11.54	287.14±11.27
50	48	332.41±12.73	291.26±10.41
50	80	334.02±10.82	292.84±9.93
50	128	332.87±13.17	292.07±11.86

As shown in Table 11, the quality threshold directly controls the fidelity of synthetic data. Low thresholds admit noisy samples that degrade manifold quality, with MADE exhibiting steeper degradation at $\tau = 0.3$ due to its compact task space imposing stricter plausibility requirements on synthetic contexts. Overly strict filtering discards valid expansion candidates and limits the effective coverage of the augmented distribution. The optimal threshold at 0.5 balances quality and quantity to maximize the effective synthetic task count.

As shown in Table 12, interpolation and extrapolation play complementary roles. Pure interpolation maintains strong IID performance but shows an OOD gap, while pure extrapolation explores boundaries at the cost of IID stability. The balanced 1:1 ratio achieves optimal performance across

Table 11: Ablation on meta-critic quality threshold on Ant-Dir.

τ_{THRESH}	AVG. FILTER RATE	IID RETURN	OOD RETURN
0.3	9.1%	327.51±16.83	284.42±15.37
0.4	22.8%	331.47±14.26	289.77±13.54
0.5	36.4%	**334.51±12.31**	**293.53±11.17**
0.6	43.2%	332.81±13.47	291.48±12.83
0.7	59.7%	329.65±15.92	287.57±14.61

Table 12: Ablation on expansion strategy composition on Ant-Dir.

STRATEGY	N_{IID} / N_{OOD}	IID RETURN	OOD RETURN
INTERPOLATION ONLY	64 / 0	333.84±13.46	285.58±14.73
EXTRAPOLATION ONLY	0 / 64	326.65±15.82	290.54±13.28
BALANCED (1:1)	32 / 32	**334.51±12.31**	**293.53±11.17**
IID-BIASED (3:1)	48 / 16	334.17±12.87	288.89±13.94
OOD-BIASED (1:3)	16 / 48	330.56±14.53	291.92±12.46

both regimes, validating that manifold densification and boundary exploration jointly minimize the distribution distance.

B.8 COMPUTATIONAL COST ANALYSIS

Table 13: Wall-clock training time comparison across five benchmarks. All experiments conducted on two NVIDIA L40S GPUs with 200k training steps. Results averaged over 4 random seeds.

METHOD	ANT-DIR	HC-DIR	HC-VEL	HOPPER-P	WALKER-P
UNICORN	8H 00M±15M	7H 47M±14M	5H 48M±11M	8H 00M±16M	6H 46M±13M
REDA	25H 52M±52M	25H 06M±47M	19H 28M±25M	25H 54M±55M	22H 22M±38M
CERTAIN	12H 34M±28M	9H 10M±22M	9H 15M±25M	12H 27M±31M	10H 18M±26M
MADE	8H 38M±17M	8H 40M±18M	6H 16M±13M	8H 38M±19M	7H 35M±16M
Overhead relative to UNICORN					
REDA	+223.3%	+221.9%	+235.6%	+223.8%	+230.5%
CERTAIN	+56.9%	+17.7%	+59.5%	+55.9%	+52.4%
MADE	+7.9%	+11.3%	+8.0%	+7.9%	+12.1%

Table 13 compares the wall-clock training time across all methods on two NVIDIA L40S GPUs. Each experiment runs for 200k training steps with identical hardware and hyperparameter configurations. MADE introduces minimal computational overhead compared to UNICORN, with an average increase of only 9.4% across all environments. This stems from the sleep phase operating on cached latent representations rather than raw trajectories, and the consolidation process executing infrequently. In contrast, CERTAIN incurs 48.5% overhead due to its uncertainty quantification mechanism requiring ensemble inference, while ReDA suffers from 227.0% overhead caused by adversarial min-max optimization with multiple discriminator updates per iteration. The modest cost increase of MADE validates its practical efficiency for offline meta-learning scenarios where the performance gains substantially outweigh the marginal training time extension.

B.9 VISUALIZATION

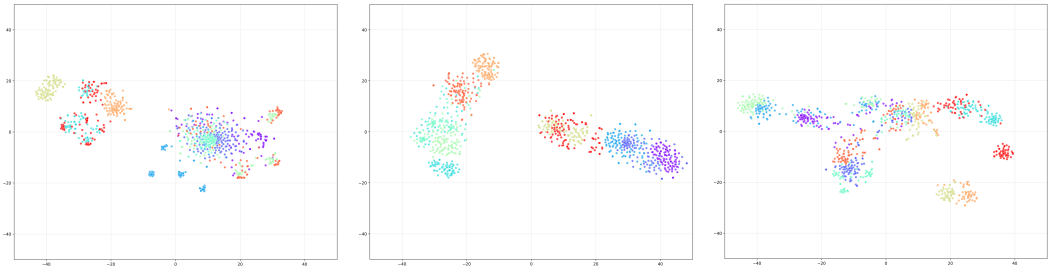


Figure 5: t-SNE visualization of task embeddings in HalfCheetah-Dir using out-of-distribution samples. Each color represents one of ten directional locomotion tasks.

Figure 5 illustrates how MADE reshapes the latent manifold compared to baselines. UNICORN produces separated task clusters with visible gaps between them, while CERTAIN shows similar fragmentation despite incorporating uncertainty estimation. These voids represent regions where encoders lack support, causing failures when OOD test tasks fall into sparse areas. MADE densifies the space through interpolation between existing prototypes and conservative extrapolation toward boundaries, yielding a more continuous manifold. The increased coverage directly explains the performance gains in Table 1, as the expanded representation space better accommodates manifold variations beyond the fixed training distribution.

Table 14: Hyperparameter configurations across five benchmark environments. The weight β corresponds to the decoder loss weight in Eq. 14, following UNICORN (Li et al., 2024).

CONFIGURATION	ANT-DIR	HALFCHEETAH-DIR	HALFCHEETAH-VEL	HOPPER-PARAM	WALKER-PARAM
Environment & Data					
DATASET SIZE	1E5	2E5	2E5	3E5	4.5E5
NUM. TRAIN TASKS	20	20	20	20	20
NUM. EVAL TASKS	20	20	20	20	20
Training Configuration					
TRAINING STEPS	200K (1000 ITERS × 200 RL UPDATES)				
CONTEXT SIZE	1 TRAJECTORY (200 STEPS)				
Network Architecture					
LEARNING RATE	3E-4				
POLICY/CRITIC WIDTH	256				
POLICY/CRITIC DEPTH	3 (2 HIDDEN LAYERS)				
ENCODER HIDDEN SIZE	64	64	64	128	128
DECODER WEIGHT β	0.15	0.01	0.15	1.5	0.7
MADE					
PAIRING TEMPERATURE τ	0.5				
INTERPOLATION BETA $(\alpha_{\text{BETA}}, \beta_{\text{BETA}})$	(2.0, 2.0)				
DIFFUSION NOISE σ^2	0.01				
EXTRAPOLATION μ_ξ/σ_ξ	0.5 / 0.2				
EXTRAPOLATION BOUNDS (a, b)	(0.0, 1.0)				
SCORE WEIGHTS $(w_{\text{CONS}}, w_{\text{DIV}}, w_{\text{CAUS}})$	(0.33, 0.33, 0.33)				
QUALITY THRESHOLD τ_{THRESH}	0.5				
REPLAY INTERVAL T_{REPLAY}	50 ITERATIONS				
IID / OOD SAMPLES $(N_{\text{IID}}, N_{\text{OOD}})$	(32, 32)				
CONSOLIDATION STEPS K_{CONS}	50				

C HYPERPARAMETER CONFIGURATIONS

Table 14 presents the hyperparameter configurations used for each benchmark environment. All experiments share common training parameters (200k training steps, task batch size 16, RL batch size 256) and network architectures (learning rate 3e-4, RL network width 256, depth 3). All methods use Soft Actor-Critic (SAC) with entropy coefficient 0.2 and discount factor 0.99. Soft target update coefficient is set to $\tau = 0.005$ for all environments. The decoder architecture follows a CVAE structure with the specified hidden size and number of layers (1 hidden layer). For MADE, the sleep phase parameters (T_{replay}, N_{IID}, N_{OOD}, K_{cons}) are kept constant across all environments to demonstrate robustness.