# Objective-Aligned Amortized Inference for Offline Bayes-Adaptive MDP Model Learning

**Toru Hishinuma**
Proxima Technology / Kyoto University
toru.hishinuma@proxima-ai-tech.com

**Kei Senda**
Kyoto University
senda@kuaero.kyoto-u.ac.jp

## Abstract

Learning latent environment representations from offline datasets is an important challenge for Bayes-adaptive Markov decision process model learning. A key difficulty is that offline datasets collected under heterogeneous protocols can differ in state–action visitation even when the underlying dynamics are identical. Visitation-decoupled objectives address this by evaluating transition predictive fit under a protocol-independent reference distribution via a change of measure. We show, however, that the resulting objective-induced invariances need not be inherited by amortized inference. An unconstrained encoder can depend on protocol-induced visitation variations that are invisible to the objective, a structural failure mode we term *amortization mismatch*. To address this issue, we formalize an *objective-aligned* design principle as an objective–inference compatibility condition: the encoder should factor through the objective-induced dataset signature. We instantiate this principle in multiple set-encoder families and, through targeted experiments, provide evidence that objective-aligned encoders improve posterior consistency and can mitigate downstream planning discrepancy across protocols.

## 1 Introduction

In many sequential decision-making settings, including model-based reinforcement learning (RL) [24] and planning/control such as model predictive control [20], environments share common structure but differ in latent properties, such as friction and mass, that affect transition dynamics. Bayes-Adaptive Markov Decision Processes (BAMDPs) [6] provide a principled framework for these settings, where the semantic role of the latent variable is to index environment-specific transition structure. In this paper, we study how to learn BAMDP models with such latent variables, focusing on offline learning.

Offline datasets are often collected under heterogeneous data-collection protocols, including behavior policies, reset methods, and logging constraints [14]. In offline datasets, state–action visitation is shaped not only by transition dynamics but also by such data-collection protocols. Even when the underlying dynamics are identical, datasets collected from the same environment instance may differ in state–action visitation. Offline datasets contain information about transition dynamics together with protocol-induced visitation variation.

For BAMDP model learning, this creates a fundamental tension: the latent variable is meant to index environment-specific transition structure, while offline datasets also contain protocol-induced visitation variation. A naive learning method can therefore entangle these two sources of variation, causing the learned latent representation to reflect properties of the data-collection protocol rather than the environment alone. From the BAMDP perspective, the intended semantics are that latent representations respond to information about transition dynamics while remaining invariant to protocol-induced visitation variation.

The desired invariance can be specified through a visitation-decoupled objective (VD objective), which evaluates transition predictive fit under a protocol-independent reference distribution over

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

state–action pairs rather than the visitation distribution, via a change of measure. From the generalized Bayes perspective [1], the choice of objective bears on the latent semantics, so it is natural to design the objective to exclude protocol-induced visitation variation from those semantics. Under the VD objective, the dataset enters the objective only through an *objective-induced dataset signature*, a weighted empirical aggregate of observed transitions. We call two datasets *objective-equivalent* if they induce the same signature, even when their raw visitation patterns differ.

What is easy to miss is that objective-induced equivalence is an objective-level notion: it captures invariances of the objective, not of the amortized inference mechanism. Crucially, it constrains what the objective can see, but it does not, by itself, constrain what an amortized encoder may see.

Objective-induced equivalence is not automatically inherited by amortized inference. The failure at issue here does not arise from finite data, encoder expressivity, or instability in estimating the weights required to implement the change of measure. Its source is structural: the VD objective sees a dataset only through the objective-induced dataset signature, whereas an unconstrained amortized encoder may still respond to dataset properties beyond that signature. As a result, even objective-equivalent datasets can yield different latent representations whenever the encoder depends on dataset features that do not factor through the signature. We refer to this mismatch between objective-induced invariance and the data dependence of amortized inference as *amortization mismatch*.

To address the amortization mismatch, we formalize an *objective-aligned* design principle for amortized inference in offline BAMDP model learning. This principle requires that the amortized inference encoder depend on an offline dataset only through the objective-induced dataset signature. The key point is that, by aligning the encoder's data dependence with that of the objective, this principle makes the VD objective's invariance inherited by amortized inference, thereby structurally preserving the intended latent semantics specified by the VD objective.

This principle is not tied to any particular encoder architecture, and we study two forms: a deepset encoder and a residual attention set encoder. Our contribution is the objective-aligned principle; the encoders are its instantiations. Using these instantiations, we present mechanism-focused experiments that isolate the structural failure mode and evaluate the design principle, rather than benchmark-driven performance comparisons. A toy construction establishes amortization mismatch and shows that objective-aligned encoders avoid it. In Pendulum and HalfCheetah, we show that objective-aligned encoders improve posterior consistency under protocol-induced visitation variation. We also examine whether the resulting posteriors lead to downstream planning discrepancy. These results suggest that the inheritance condition is structural across architectures and beneficial when enforced.

Our main contributions are summarized as follows.

- We identify *amortization mismatch*, a structural failure in which unconstrained encoders can violate the dataset equivalence induced by the VD objective for BAMDP latent semantics.
- We formalize an *objective-aligned* design principle as an objective–inference compatibility condition: the encoder should factor through the objective-induced dataset signature, so that amortized inference inherits the objective-induced invariance.
- We instantiate the principle in multiple set-encoder families and present empirical evidence that objective-aligned encoders improve posterior consistency and can mitigate downstream planning discrepancy under protocol-induced visitation variation.

## 2 Related Works

**Map and claim.** We connect three ingredients that are often discussed separately: (R1) change-of-measure treatments of covariate shift; (R2) the dataset dependence induced by architectural choices for amortized inference; and (R3) representation/context shift in offline/meta-RL under behavior-policy mismatch and training effects. Our contribution lies in linking (R1) objective-level shift correction to (R2) the dataset dependence of amortized encoders. Specifically, even when (R1) succeeds, (R2) can fail unless the encoder is explicitly aligned to the objective-induced dataset signature. We refer to (R3) for context: stability-focused shift-mitigation approaches are complementary, but do not by themselves ensure inheritance of objective-induced invariances by amortized inference. Below, we begin with background, and then discuss (R1)–(R3) in turn.

**Background: latent environments and offline data.** A common modeling device augments an MDP with an environment or task variable that captures transition variation across conditions. In contextual MDPs this variable is observed [16], while latent-variable MDPs infer it from experience [12]. BAMDPs similarly treat unknown environment descriptors, such as transition-model parameters, as latent variables governing transition dynamics and formalize decision-making under posterior uncertainty [4, 6], motivating Bayes-adaptive planning [7]. Meta-RL operationalizes this idea via amortized posteriors over latent task variables inferred from experience and used to condition downstream components [19, 34]. Recent offline meta-RL pipelines couple inferred task representations with learned models (e.g., contextual world models) to support generalization from offline logs [28]. In offline meta-RL, learning such beliefs from fixed logs raises identifiability and ambiguity issues, where multiple latent explanations can fit the same logs [3]. More broadly in offline RL, protocol-dependent variation in state–action visitation is unavoidable, motivating conservative decision-making criteria to mitigate out-of-distribution error [11, 30]. Our work shares two threads—latent inference and protocol-induced shift—but targets a complementary issue: we focus on protocol sensitivity of latent semantics, rather than conservative decision making or offline identifiability/ambiguity.

**(R1) Objective semantics: importance weighting and criterion-defined semantics.** Importance weighting corrects covariate shift by reweighting samples to match a reference distribution [21, 22]. In off-policy evaluation, change-of-measure methods estimate target-policy returns by using a policy-induced distribution as a reference; the required importance weights can be obtained by approaches such as [17]. In offline model-based RL, importance weighting can similarly use a target-policy-induced distribution as the reference to fit dynamics so as to improve accuracy in regions of interest [29], including in variational BAMDP model learning [9]. In contrast, we use change-of-measure to define semantics: we choose a change-of-measure objective to prevent the semantics from being shaped by visitation variation, under a user-specified reference distribution, possibly induced by a policy. This criterion-defined view is supported by loss-based Bayesian updating [1], which formalizes how a chosen objective endows inferred quantities with an interpretation—that is, their semantics under the criterion.

**(R2) Amortized inference from transition sets: architectural invariances and dataset dependence.** Amortized inference of latent task/environment variables from transition sets is often implemented with permutation-invariant set encoders [32, 13]. These encoders are standard in meta-RL pipelines that infer a latent task from experience and condition downstream components [19]. However, permutation invariance alone does not fix dataset dependence: pooling/attention can still react to objective-invisible features. Our contribution is to align the encoder's dataset dependence with that of the objective: under a VD objective, the encoder should depend on the dataset only through the objective-induced dataset signature.

**(R3) Amortized inference under shift: representation/context shift versus amortization mismatch.** Prior work in context-based offline meta-RL has studied stability-focused methods for mitigating representation or context shift induced by behavior-policy mismatch or training dynamics. Examples include CORRO [31], CSRO [5], analyses of task representation shift [33], and UNICORN [15]. These works primarily study task representations used by policy-side components, such as meta-policies, actor-critic objectives, or adaptation procedures, rather than representations that serve as environment indices of an explicit transition model. Our focus is not stability-focused shift mitigation, but characterizing when protocol-induced visitation variation is excluded from the learned latent environment index of a transition model. Reducing representation or context drift alone does not ensure that the learned latent is invariant to protocol-induced visitation variation. We formalize this requirement as objective-invariance: datasets with the same objective-induced signature should induce the same latent belief. We refer to violations of this requirement as amortization mismatch, which is orthogonal to the amortization gap in variational inference [2].

## 3 Preliminaries

### 3.1 Offline Latent Environment Setting

We consider a family of environments sharing state and action spaces $(\mathcal{S}, \mathcal{A})$. Let $s$ denote a state and $a$ an action. Let $\xi$ denote an unobserved environment descriptor that governs the true transition

dynamics, $p^\star(s' \mid s, a; \xi)$. We assume that $\xi$ is not observable to the agent and admits no known parameterization; moreover the functional form of $p^\star(\cdot \mid \cdot, \cdot; \xi)$ is unknown. For simplicity we focus on latent variation in transitions; extending to environment-dependent rewards is straightforward.

We introduce a latent variable $\theta \in \Theta$, inferred from data and used to condition a transition model. The latent $\theta$ is not assumed to coincide with $\xi$ or any physical parameter; its semantics are determined by the learning objective and the inference architecture.

We are given offline datasets from $M$ environments: $\mathcal{D}_m := \{ (s_{m,i}, a_{m,i}, s'_{m,i}) \}_{i=1}^{N_m}$, $\mathcal{D} := \{\mathcal{D}_m\}_{m=1}^{M}$. Let $\xi_m$ be the (unobserved) environment descriptor of the $m$-th environment, and let $\theta_m \in \Theta$ be the environment-specific latent inferred from $\mathcal{D}_m$. Let $\mu_m(s, a)$ denote the (unknown) state–action distribution induced by the data-collection protocol in the $m$-th environment; it need not be stationary and is generally not determined by $\xi_m$ alone.

Let $\mu_{\mathrm{obs},m}(s, a)$ be the empirical state–action visitation distribution of $\mathcal{D}_m$. We neither model nor estimate $\mu_{\mathrm{obs},m}$; it is used only to motivate a change-of-measure view of visitation decoupling.

**BAMDP perspective.** In a BAMDP, the environment descriptor $\xi$ captures intrinsic transition properties and is, by definition, independent of data-collection protocol choices (behavior policies, reset methods, etc.). Since $\xi$ is unobserved, we use $\theta$ as a learned environment index. Therefore, protocol variations that change visitation without changing the true dynamics should ideally not be reflected in the inferred latent representation.

### 3.2 Latent-Parameter Transition Model

We place a prior $p(\theta)$ over latent variables and learn a conditional transition model. For the $m$-th environment, we draw $\theta_m \sim p(\theta)$ and model transitions as $s' \sim p_\psi(s' \mid s, a, \theta_m)$. The model $p_\psi$ is trained to match environment-specific transitions in offline data. We make no assumption that $\theta_m$ corresponds to the true environment descriptor $\xi_m$; rather, $\theta_m$ is a learned representation whose meaning is induced by the objective and inference procedure.

A standard likelihood-based approach considers maximizing the log-marginal likelihood objective,

$$\log p_\psi(\mathcal{D}) = \sum_{m=1}^{M} \log \int p(\theta_m) \prod_{i=1}^{N_m} p_\psi(s'_{m,i} \mid s_{m,i}, a_{m,i}, \theta_m)\, d\theta_m. \tag{1}$$

## 4 Objective-Aligned Amortized Inference

### 4.1 Visitation Variation and BAMDP Latent Semantics

Offline datasets contain information about environment-dependent transition dynamics together with protocol-induced visitation variation arising from data-collection protocols. The standard objective (1) does not separate these two sources of variation: because it evaluates transition predictive fit under the dataset's empirical visitation, it implicitly defines a *protocol-dependent* semantics. Under this semantics, the learned latent representation can come to reflect protocol-induced visitation variation in addition to environment-specific transition structure. From the BAMDP perspective, this is undesirable: the latent variable is intended to serve as an environment index for transition dynamics, not as a representation of protocol variation. This motivates defining a *protocol-independent* semantics under which protocol-induced visitation variation does not enter the learned latent semantics.

### 4.2 Visitation-Decoupled Objective and Dataset Signature

To define a protocol-independent semantics, we replace dataset-visitation evaluation with protocol-independent reference evaluation, implemented below via importance weighting. Let $\rho(s, a)$ be a reference distribution that specifies the semantics—i.e., the state–action regions on which protocols are compared—and evaluate predictive fit under $\rho$ rather than the dataset visitation.

For the $m$-th environment, we define the visitation-decoupled semantics by maximizing the reference-evaluated expected log-likelihood of one-step transitions,

$$\mathbb{E}_{(s,a)\sim\rho}[\log p_\psi(s' \mid s, a, \theta_m)]\,. \tag{2}$$

Since transitions in $\mathcal{D}_m$ are observed under the data-collection $\mu_m(s,a)$, we estimate (2) via a change of measure with importance weights $w_{m,i} = \frac{\rho(s_{m,i},a_{m,i})}{\mu_m(s_{m,i},a_{m,i})}$. Using $\mathcal{D}_m$, we obtain the following empirical estimator of (2) over $N_m$ transitions:

$$L(\psi, \theta_m; \mathcal{D}_m) = \sum_{i=1}^{N_m} w_{m,i} \, \log p_\psi(s'_{m,i} \mid s_{m,i}, a_{m,i}, \theta_m), \tag{3}$$

which we interpret as a Monte Carlo change-of-measure approximation of (2). We assume $\rho << \mu_m$ (or use stabilized surrogates otherwise) so that $w_{m,i}$ is well-defined.

Analogous to the standard objective (1), we define the VD objective as

$$\log \tilde{p}_\psi(\mathcal{D}; \rho) := \sum_{m=1}^{M} \log \int p(\theta_m) \, e^{L(\psi, \theta_m; \mathcal{D}_m)} \, d\theta_m, \tag{4}$$

where the change of measure is encoded in $L(\psi, \theta_m; \mathcal{D}_m)$ through the importance weights. The VD objective (4) evaluates transition predictive fit under the protocol-independent reference distribution $\rho$ rather than under the dataset's empirical visitation. We therefore interpret (4) as defining a protocol-independent semantics, consistent with the generalized-Bayes view that the chosen criterion determines the resulting semantics [1].

**Dataset signature.** Define the weighted empirical aggregates

$$\nu(\mathcal{D}_m) := \sum_{i=1}^{N_m} w_{m,i} \, \delta_{(s_{m,i}, a_{m,i}, s'_{m,i})}. \tag{5}$$

For any fixed $\theta_m$, the weighted log-likelihood (3) depends on $\mathcal{D}_m$ only through $\nu(\mathcal{D}_m)$. We refer to $\nu(\mathcal{D}_m)$ as the *objective-induced dataset signature*.

**Definition (Objective-induced equivalence).** Two datasets $\mathcal{D}_m$ and $\mathcal{D}'_m$ are *objective-equivalent*, written $\mathcal{D}_m \sim \mathcal{D}'_m$, if $\mathcal{D}_m \sim \mathcal{D}'_m \iff \nu(\mathcal{D}_m) = \nu(\mathcal{D}'_m)$.

**Observation (Objective-induced invariance).** If $\mathcal{D}_m \sim \mathcal{D}'_m$, then for any fixed $\theta_m$, $L(\psi, \theta_m; \mathcal{D}_m) = L(\psi, \theta_m; \mathcal{D}'_m)$. Hence, replacing $\mathcal{D}_m$ by an objective-equivalent dataset does not change the VD objective (4).

### 4.3 Visitation-Decoupled ELBO

To maximize the VD objective (4), we introduce a variational posterior $q_\phi(\theta_m \mid \mathcal{D}_m)$ and optimize the resulting VAE-style criterion (cf. [10]). Following $\beta$-VAE [8], we rescale the KL term by $\beta$.

$$\mathcal{L}_{\mathrm{VD}}(\psi, \phi; \mathcal{D}) = \sum_m \frac{1}{Z_m} \left\{ \mathbb{E}_{\theta_m \sim q_\phi(\cdot \mid \mathcal{D}_m)} \Big[ L(\psi, \theta_m; \mathcal{D}_m) \Big] - \beta \mathrm{KL}(q_\phi(\theta_m \mid \mathcal{D}_m) \,\|\, p(\theta_m)) \right\}, \tag{6}$$

where $w_{m,i}$ are importance weights and $Z_m := \sum_{i=1}^{N_m} w_{m,i}$. We call this $\beta$-weighted criterion the VD-ELBO, without implying a strict ELBO lower-bound for $\beta \neq 1$.

### 4.4 Amortization Mismatch

A natural desideratum is that latent inference should respect the same invariance: $\mathcal{D}_m \sim \mathcal{D}'_m$ should imply $q_\phi(\theta_m \mid \mathcal{D}_m) = q_\phi(\theta_m \mid \mathcal{D}'_m)$. However, "$\sim$" only captures invariances of the objective-induced equivalence classes defined by $\nu(\cdot)$. It does not restrict an amortized encoder, which can vary within an equivalence class and map $\mathcal{D}_m$ and $\mathcal{D}'_m$ to different posteriors despite $\mathcal{D}_m \sim \mathcal{D}'_m$.

**Lemma 1 (Counterexample: non-inheritance under objective-equivalent datasets).** There exist two objective-equivalent datasets $\mathcal{D}_m \sim \mathcal{D}'_m$, a reference distribution $\rho$, and an unconstrained amortized encoder $q_\phi(\theta_m \mid \mathcal{D}_m)$ such that $q_\phi(\theta_m \mid \mathcal{D}_m) \neq q_\phi(\theta_m \mid \mathcal{D}'_m)$.

**Proof sketch.** Construct $\mathcal{D}'_m$ from $\mathcal{D}_m$ by adding a transition whose importance weight under $\rho$ is zero. This leaves the objective-induced dataset signature unchanged, $\nu(\mathcal{D}_m) = \nu(\mathcal{D}'_m)$, so $\mathcal{D}_m \sim \mathcal{D}'_m$. A deepset encoder with unweighted pooling, however, aggregates the transition set without weighting and can therefore produce different posteriors for $\mathcal{D}_m$ and $\mathcal{D}'_m$. A formal proof is given in Appendix A.1; Sec. 5.1 gives the corresponding constructive illustration.

**Definition (Amortization mismatch).** An amortized inference model $q_\phi(\theta_m \mid \mathcal{D}_m)$ exhibits *amortization mismatch* if there exist $\mathcal{D}_m, \mathcal{D}'_m$ with $\nu(\mathcal{D}_m) = \nu(\mathcal{D}'_m)$ such that

$$q_\phi(\theta_m \mid \mathcal{D}_m) \neq q_\phi(\theta_m \mid \mathcal{D}'_m). \tag{7}$$

This motivates enforcing objective-induced invariances at the encoder level.

### 4.5 Design Principle: Objective-Aligned Amortized Inference

This invariance is specified by objective-induced equivalence: replacing $\mathcal{D}_m$ by an objective-equivalent dataset leaves the VD objective unchanged. To align an amortized encoder with the semantics specified by the VD objective, we require the encoder to respect the same equivalence.

**Requirement (Objective-invariant inference).** We call an amortized posterior $q_\phi(\theta_m \mid \mathcal{D}_m)$ *objective-invariant* if $\mathcal{D}_m \sim \mathcal{D}'_m \implies q_\phi(\theta_m \mid \mathcal{D}_m) = q_\phi(\theta_m \mid \mathcal{D}'_m)$.

**Objective-aligned design principle.** We operationalize this requirement by *restricting the encoder's data dependence* to the objective-induced dataset signature:

$$q_\phi(\theta_m \mid \mathcal{D}_m) = Q_\phi\big(\nu(\mathcal{D}_m)\big), \tag{8}$$

where $Q_\phi$ is a function of $\nu(\mathcal{D}_m)$ defined in (5). We refer to this parameterization as *objective-aligned amortized inference*.

With this parameterization, the encoder and the VD-ELBO share the same objective-induced invariance, since each environment-wise term depends on $\mathcal{D}_m$ only through $\nu(\mathcal{D}_m)$.

**Remark.** Once the relevant equivalence relation is identified, the factorization condition follows directly; see Appendix A.2. The key point is to identify the relation: (i) because a BAMDP latent indexes transition dynamics, its semantics require a criterion that excludes protocol-induced visitation variation; (ii) the VD objective provides this criterion and induces a dataset equivalence relation; (iii) objective alignment then enforces this equivalence through the encoder's data dependence.

### 4.6 Simple Instantiations: Objective-Aligned Set Encoders

We show how the objective-aligned principle above is realized in set encoders. Many set-encoder families can be made objective-aligned by redefining their raw set aggregation as aggregation under the objective-induced weighted empirical measure. We illustrate this idea with a deepset encoder and a residual attention set encoder.

**Objective-aligned deepset encoder.** Let $g_\phi$ be an embedding function, and define the weighted set

$$h(\mathcal{D}_m) = \sum_{i=1}^{N_m} w_{m,i}\, g_\phi(s_{m,i}, a_{m,i}, s'_{m,i}). \tag{9}$$

This aggregation is permutation-invariant, and because it uses the same weights as the objective, $h(\mathcal{D}_m)$ is a function of $\nu(\mathcal{D}_m)$. We then parameterize the variational posterior as

$$q_\phi(\theta_m \mid \mathcal{D}_m) = \mathcal{N}\Big(\mu_\phi(h(\mathcal{D}_m), Z_m), \mathrm{diag}(\sigma_\phi(h(\mathcal{D}_m), Z_m)^2)\Big). \tag{10}$$

By construction, the encoder factors through $\nu(\mathcal{D}_m)$ and satisfies the objective-aligned principle (8).

**Objective-aligned residual attention set encoder.** Objective alignment is not limited to deepset-style set-encoding. We also consider a residual attention set encoder that augments a weighted global summary with a scaled residual refinement computed from centered element features and aggregated under the same weight measure. This allows the encoder to incorporate limited element-wise interactions while keeping its dataset dependence restricted to $\nu(\mathcal{D}_m)$. Hence, this encoder also satisfies the objective-aligned principle (8). Implementation details are in Appendix B.3.

**Remark.** These instantiations illustrate that objective alignment is an architecture-agnostic principle that constrains the encoder's dataset dependence. The particular set-encoder architecture may vary, but the essential structural constraint is that the encoder factors through $\nu(\mathcal{D}_m)$. Equivalently, this constraint makes the amortized posterior constant on objective-induced equivalence classes.

### 4.7 Weight Interpolation for Practical Weight Construction Noise

When accurate weights are available, the VD-ELBO with an objective-aligned encoder provides a direct construction of the intended protocol-independent latent semantics. In practice, however, weights often need to be constructed from data and then contain weight-construction noise. Using the raw weights is, in principle, the most direct way for the intended latent semantics, but can be unstable in practice when the weights are substantially noisy or heavy-tailed. Conversely, ignoring the weights remains sensitive to protocol-induced visitation variation but is stable against weight noise.

We consider weight interpolation between the unweighted endpoint and the raw-weight endpoint as a way to stabilize the shared weighting while approximately preserving the intended latent semantics. This interpolation can be viewed as an instance of a widely used practical weighting idea, and we apply it to the empirical weighting shared by the VD-ELBO and the objective-aligned encoder. As one simple weighting choice, Section 5.3 reports an interpolation study based on tempering, $(w_{m,i})^{\alpha}$.

## 5 Experiments

### 5.1 Toy: Objective-Induced Invariance Does Not Inherit "for Free"

This subsection gives a constructive counterexample for Lemma 1 by making amortization mismatch explicit, and shows how the objective-aligned principle restores the intended invariance.

We use a single-environment toy BAMDP with linear-Gaussian dynamics to isolate the phenomenon in (7); details are given in Appendix C. Let $\mathcal{D}_1$ contain transitions from a relevant state–action region $\mathcal{R}$. For each $N \geq 0$, we construct $\mathcal{D}'_{1,N} = \mathcal{D}_1 \cup \mathcal{I}_N$, where $\mathcal{I}_N$ consists of $N$ independently sampled transitions from an irrelevant region $\mathcal{I}$. Since transitions from $\mathcal{I}$ receive zero importance weight, $\nu(\mathcal{D}_1) = \nu(\mathcal{D}'_{1,N})$ for every $N$. In particular, $\mathcal{D}'_{1,0} = \mathcal{D}_1$.

We study two encoder architectures, the deepset and the residual attention set in Sec. 4.6. For each architecture, we instantiate both an unconstrained encoder and an objective-aligned encoder. For each encoder, we train $q_\phi$ once on $\mathcal{D}_1$, then evaluate the same fixed $q_\phi$ on $\mathcal{D}_1$ and $\mathcal{D}'_{1,N}$. We measure posterior sensitivity by the squared Wasserstein-2 distance $W_2^2\big(q_\phi(\theta_1 \mid \mathcal{D}_1),\, q_\phi(\theta_1 \mid \mathcal{D}'_{1,N})\big)$ as $N$ varies.

Figure 1 instantiates Lemma 1 for both encoder architectures. Although the datasets remain objective-equivalent for every $N$, unconstrained encoders exhibit increasing posterior discrepancy, whereas objective-aligned encoders remain invariant up to numerical precision. The similar behavior across the two architectures suggests that the need for alignment is structural rather than specific to a particular encoder family.

### 5.2 Pendulum: Posterior Consistency and Downstream Planning Discrepancy

This subsection evaluates whether protocol-induced visitation variation induces posterior inconsistency across protocols, and whether objective-aligned inference improves posterior consistency. We also examine whether differences in the resulting posteriors lead to downstream planning discrepancy.

We use PENDULUM-V1, where the environment descriptor $\xi$ is the pendulum length. For each $\xi$, we collect two datasets that differ only in the reset distribution: **Protocol A** uses truncated-Gaussian resets, whereas **Protocol B** uses uniform resets. Both protocols use the same random-action policy. We use these protocols only to manipulate visitation, not to model realistic logging pipelines. We consider $\xi \in \{0.7, 1.0, 1.3\}$ and two protocols, yielding $M{=}6$ datasets, indexed by $m$ with $(\xi_m, \mathrm{protocol}_m)$ denoting the corresponding pair. We set $\rho$ to the empirical mixture over these datasets, giving a neutral, non-privileging reference.

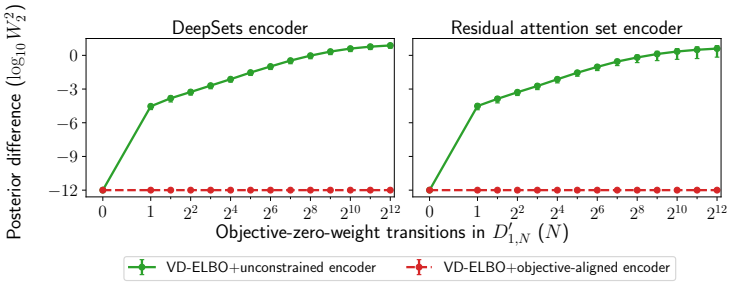


Figure 1: **Toy (Lemma 1): amortization-mismatch test under exact objective-equivalent datasets.** For each $N$, we construct $\mathcal{D}'_{1,N} = \mathcal{D}_1 \cup \mathcal{I}_N$ by adding $N$ objective-zero-weight transitions to $\mathcal{D}_1$, so that $\nu(\mathcal{D}_1) = \nu(\mathcal{D}'_{1,N})$ and $\mathcal{D}'_{1,0} = \mathcal{D}_1$. We fix $q_\phi$ trained on $\mathcal{D}_1$ and plot $\log_{10} W_2^2\big(q_\phi(\theta_1 \mid \mathcal{D}_1), q_\phi(\theta_1 \mid \mathcal{D}'_{1,N})\big)$. In both deepset and residual attention set encoder architectures, unconstrained encoders are sensitive to these transitions, while objective-aligned encoders remain invariant up to numerical precision. Values below $10^{-12}$ are displayed at the plotting floor. Error bars denote $\pm 1$ standard error across 5 seeds, computed on $W_2^2$ before log transformation.

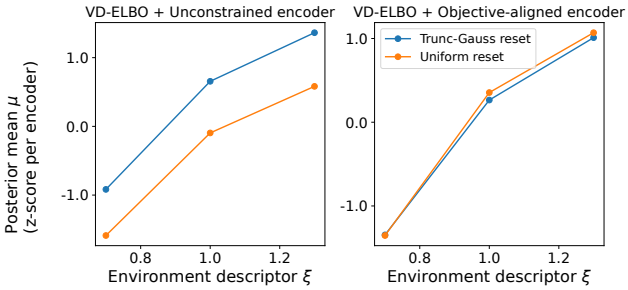


Figure 2: **Pendulum: posterior tracking (deepset, estimated weights, $h$=1, seed 0).** The unconstrained encoder exhibits protocol-dependent drift at fixed $\xi$, whereas the objective-aligned encoder largely removes this drift while preserving $\xi$-sensitivity.

We use a context-subsampled minibatch surrogate for training (Appendix B.2). We report results for data-collecting episode horizon $h$=1 with estimated weights in Sec. 5.2. Oracle weights, $h$>1 results, and additional diagnostics are provided in Appendix D.

**Posterior Consistency.** After training, for each $(\xi, \text{protocol})$, we compute the full-data posterior and compare how the posterior mean varies with $\xi_m$ across protocols. Figure 2 shows that the unconstrained deepset encoder exhibits protocol-dependent offsets at fixed $\xi$, whereas the objective-aligned deepset encoder largely aligns the two protocol curves while preserving $\xi$-sensitivity. Across five seeds, objective alignment reduces relative protocol drift from $0.265$ to $0.020$ for deepset and from $0.141$ to $0.023$ for residual attention, while preserving $\xi$-sensitivity. Appendix D provides the metric definitions, seed-wise plots, and full quantitative summaries for both encoders.

**Downstream Planning Discrepancy.** After training, for each $(\xi, \text{protocol})$, we compute the full-data posterior, plan open-loop actions with a fixed CEM planner in the learned model, and evaluate returns in the true environment. Table 1 reports the planning discrepancy, i.e., the difference in achieved return between the protocols at the same $\xi$. Because downstream discrepancy is not determined by the posterior alone, we use this evaluation as a diagnostic of whether reducing protocol-dependent posterior drift can carry over to planning. We also include an unconstrained encoder trained with a standard ELBO as a naive method that ignores visitation-decoupling in both the objective and the encoder. The results indicate that posterior inconsistency can lead to downstream planning discrepancy, and that objective-aligned encoders can mitigate it. The gap change can be less pronounced in noisier settings with less smooth dynamics, see Appendix E.5.

Table 1: **Pendulum: downstream planning discrepancy (estimated weights, $h$=1; mean $\pm$ std across 5 seeds).** Posterior inconsistency can lead to planning discrepancy, and objective-aligned encoders (OAEs) can mitigate this discrepancy, compared to unconstrained encoders (UEs).

| Method | $\xi = 0.7$ | $\xi = 1.0$ | $\xi = 1.3$ | mean over $\xi$ |
|---|---|---|---|---|
| deepset: ELBO+UE | $16.418 \pm 18.506$ | $0.615 \pm 0.742$ | $0.228 \pm 0.412$ | $5.754 \pm 6.069$ |
| deepset: VD-ELBO+UE | $5.611 \pm 4.492$ | $0.467 \pm 0.784$ | $0.404 \pm 0.372$ | $2.160 \pm 1.340$ |
| deepset: VD-ELBO+OAE | $0.795 \pm 0.538$ | $0.072 \pm 0.068$ | $0.096 \pm 0.075$ | $\mathbf{0.320 \pm 0.170}$ |
| res-attn: ELBO+UE | $4.608 \pm 5.673$ | $0.645 \pm 0.479$ | $0.252 \pm 0.171$ | $1.835 \pm 1.868$ |
| res-attn: VD-ELBO+UE | $1.341 \pm 0.895$ | $0.944 \pm 0.863$ | $0.232 \pm 0.273$ | $0.839 \pm 0.637$ |
| res-attn: VD-ELBO+OAE | $0.217 \pm 0.128$ | $0.220 \pm 0.149$ | $0.037 \pm 0.028$ | $\mathbf{0.158 \pm 0.073}$ |

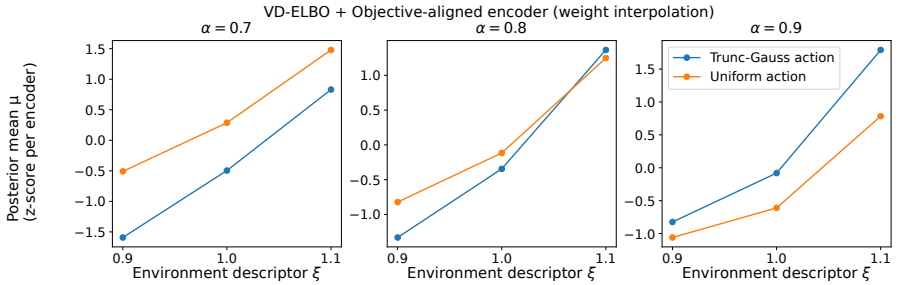


Figure 3: **HalfCheetah: posterior tracking (deepset, estimated weights, $h$=100, seed 0).** The intermediate value of weight interpolation tempering, $\alpha = 0.8$, exhibits relatively stable alignment between $\mu$ and $\xi$ under the objective-aligned encoder.

### 5.3 HalfCheetah: Weight Interpolation for Approximate Objective Alignment

This subsection examines the behavior of the weight interpolation through posterior tracking.

We use HALFCHEETAH-V5, where the environment descriptor $\xi$ is the damping parameter. Figure 3 compares posterior tracking when varying the tempering coefficient $\alpha$, showing that the intermediate value, $\alpha = 0.8$, is relatively stable. This suggests that weight interpolation can stabilize the shared weighting while approximately preserving the intended alignment. Appendix E provides experimental details, quantitative summaries, seed-wise plots across $\alpha$, and downstream planning diagnostics.

## 6 Limitations

Our contribution, the objective-aligned principle, is theoretically orthogonal to weight construction. In practice, however, the quality of the resulting latent inference can be affected by both the choice of the reference distribution $\rho$ and the quality of the corresponding weights. Choosing $\rho$ to favor well-covered state-action regions can mitigate support mismatch but may exclude informative transition differences, whereas emphasizing poorly covered regions can make importance weighting unreliable. These difficulties motivate both reference design that accounts for data coverage and improved weight construction, including post-hoc stabilization such as PSIS [27] and balance-based approaches such as stable balancing weights [35], which construct weights by balancing covariates rather than directly estimating density ratios. The results in Sec. 5.3 suggest that stabilized weights can still yield approximate alignment. Designing reference distributions to emphasize downstream-relevant regions, potentially by adapting ideas from alternating model-policy learning [29], is also promising.

Our downstream evaluations diagnose how posterior inconsistency can propagate to open-loop planning discrepancy under a fixed planner, but they do not fully characterize closed-loop behavior, where posterior inconsistency is coupled through feedback with transition-modeling error and downstream decision-making variability, such as planning or RL policy learning. Understanding whether, when, and how such coupling affects end-to-end returns remains future work.

## 7 Conclusion

We identify a structural failure mode, *amortization mismatch*, and formalize an *objective-aligned* design principle. We provide targeted evidence to isolate the mechanism: the toy construction establishes the phenomenon exactly, the Pendulum shows it in posterior tracking and downstream planning diagnostics, and the HalfCheetah suggests that it approximately remains via weight interpolation. End-to-end BAMDP pipelines involve multiple components, including model prediction, latent inference, planning or policy learning, and interaction with the true environment. By advancing the latent-inference component, the objective-aligned principle provides a mechanistic basis for the future development of end-to-end BAMDP pipelines in more practical settings.

## Acknowledgments and Disclosure of Funding

We thank all anonymous reviewers for their constructive comments.

**Funding.** This work received no dedicated research funding. A laptop and a ChatGPT subscription provided by Proxima Technology were used for research and manuscript preparation.

**Competing Interests.** The authors declare no additional competing interests beyond the affiliations and support disclosed above.

## References

[1] Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[2] Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, pages 1078–1086, 2018.

[3] Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta reinforcement learning – identifiability challenges and effective data collection strategies. In *Neural Information Processing Systems*, 2021.

[4] Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst, 2002.

[5] Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, Qi Guo, Ling Li, and Yunji Chen. Context shift reduction for offline meta-reinforcement learning. In *Neural Information Processing Systems*, volume 36, pages 80024–80043, 2023.

[6] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5–6):359–483, 2015.

[7] Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Neural Information Processing Systems*, 2012.

[8] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. $\beta$-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[9] Toru Hishinuma and Kei Senda. Importance-weighted variational inference model estimation for offline Bayesian model-based reinforcement learning. *IEEE Access*, 11:145579–145590, 2023.

[10] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[11] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.

[12] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for latent MDPs: Regret guarantees and a lower bound. In *Neural Information Processing Systems*, 2021.

[13] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, volume 97, pages 3744–3753, 2019.

[14] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

[15] 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. In *Neural Information Processing Systems*, pages 75642–75667, 2024.

[16] Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, volume 83, pages 597–618, 2018.

[17] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Neural Information Processing Systems*, volume 32, 2019.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[19] 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*, volume 97, pages 5331–5340, 2019.

[20] James Blake Rawlings, David Q Mayne, and Moritz Diehl. *Model predictive control: theory, computation, and design*. Madison, WI: Nob Hill Publishing, 2 edition, 2017.

[21] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[22] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007.

[23] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[24] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

[25] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

[26] Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo De Lazcano Perez-Vicente, J K Terry, Andrea Pierré, Sander V Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[27] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

[28] Zhi Wang, Li Zhang, Wenhao Wu, Yuanheng Zhu, Dongbin Zhao, and Chunlin Chen. Meta-DT: Offline meta-rl as conditional sequence modeling with world model disentanglement. In *Neural Information Processing Systems*, volume 37, pages 44845–44870, 2024.

[29] Shentao Yang, Shujian Zhang, Yihao Feng, and Mingyuan Zhou. A unified framework for alternating offline model training and policy learning. In *Neural Information Processing Systems*, 2022.

[30] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Neural Information Processing Systems*, 2020.

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

[32] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. In *Neural Information Processing Systems*, 2017.

[33] Hai Zhang, Boyuan Zheng, Tianying Ji, JinHang Liu, Anqi Guo, Junqiao Zhao, and Lanqing Li. Scrutinize what we ignore: Reining in task representation shift of context-based offline meta reinforcement learning. In *International Conference on Learning Representations*, 2025.

[34] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020.

[35] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

# A Formal Results on Amortization Mismatch and Objective-Invariance

## A.1 Proof of Lemma 1

We give an explicit construction for a single environment $m$. Consider two datasets

$$\mathcal{D}_m = \{(s_{m,1}, a_{m,1}, s'_{m,1})\},$$

and

$$\mathcal{D}'_m = \{(s_{m,1}, a_{m,1}, s'_{m,1}), (s_{m,2}, a_{m,2}, s'_{m,2})\}.$$

Choose a reference distribution $\rho$ such that the importance weights in the VD objective satisfy $w_{m,1} = 1$ and $w_{m,2} = 0$. For example, this holds if $\rho(s_{m,1}, a_{m,1}) = \mu_m(s_{m,1}, a_{m,1})$ and $\rho(s_{m,2}, a_{m,2}) = 0$, with $\mu_m(s_{m,2}, a_{m,2}) > 0$.

By the definition of the objective-induced dataset signature,

$$\nu(\mathcal{D}_m) = w_{m,1}\delta_{(s_{m,1},a_{m,1},s'_{m,1})} = \delta_{(s_{m,1},a_{m,1},s'_{m,1})}.$$

Similarly,

$$\nu(\mathcal{D}'_m) = w_{m,1}\delta_{(s_{m,1},a_{m,1},s'_{m,1})} + w_{m,2}\delta_{(s_{m,2},a_{m,2},s'_{m,2})}.$$

Since $w_{m,1} = 1$ and $w_{m,2} = 0$, this gives

$$\nu(\mathcal{D}'_m) = \delta_{(s_{m,1},a_{m,1},s'_{m,1})} + 0 \cdot \delta_{(s_{m,2},a_{m,2},s'_{m,2})} = \delta_{(s_{m,1},a_{m,1},s'_{m,1})}.$$

Therefore $\nu(\mathcal{D}_m) = \nu(\mathcal{D}'_m)$, and hence $\mathcal{D}_m \sim \mathcal{D}'_m$, namely, the two datasets are objective-equivalent.

It remains to construct an unconstrained amortized encoder that maps these two objective-equivalent datasets to different posteriors. Let the encoder be a deepset encoder with unweighted pooling,

$$h_\phi(\mathcal{D}) = \sum_{(s,a,s')\in\mathcal{D}} g_\phi(s, a, s'),$$

where $g_\phi$ is chosen so that $g_\phi(s_{m,1}, a_{m,1}, s'_{m,1}) = 0$ and $g_\phi(s_{m,2}, a_{m,2}, s'_{m,2}) = 1$. Let the amortized posterior be parameterized as $q_\phi(\theta_m \mid \mathcal{D}) = \mathcal{N}(h_\phi(\mathcal{D}), 1)$. Then $h_\phi(\mathcal{D}_m) = 0$ and $h_\phi(\mathcal{D}'_m) = 1$. Therefore $q_\phi(\theta_m \mid \mathcal{D}_m) = \mathcal{N}(0, 1)$ and $q_\phi(\theta_m \mid \mathcal{D}'_m) = \mathcal{N}(1, 1)$. Consequently, $q_\phi(\theta_m \mid \mathcal{D}_m) \neq q_\phi(\theta_m \mid \mathcal{D}'_m)$, even though $\mathcal{D}_m \sim \mathcal{D}'_m$. This proves the claim.

## A.2 Objective-Invariance as Factorization through the Dataset Signature

We make explicit the inheritance condition used in Sec. 4. The key point is that the equivalence relation is induced by the learning objective, whereas an amortized encoder is a separate map from datasets to posteriors. Therefore, objective-level invariance does not automatically become an invariance of amortized inference.

Once the objective-induced dataset signature $\nu(D_m)$ is identified, the objective-to-inference compatibility condition has a direct necessary and sufficient form: an amortized posterior inherits the objective-induced invariance if and only if its dependence on the dataset factors through $\nu(D_m)$.

Let $\mathcal{D}$ denote the space of offline datasets for one environment, and let $F_\phi(D_m) := q_\phi(\theta_m \mid D_m)$ denote the amortized posterior returned by an encoder. Recall that objective-induced equivalence is defined by $D_m \sim D'_m \iff \nu(D_m) = \nu(D'_m)$. Let $\mathcal{V}_\nu := \{\nu(D_m) : D_m \in \mathcal{D}\}$ denote the range of dataset signatures.

**Lemma 2.** An amortized encoder $F_\phi$ is objective-invariant, i.e., $\nu(D_m) = \nu(D'_m) \implies F_\phi(D_m) = F_\phi(D'_m)$, if and only if there exists a map $Q_\phi$ defined on $\mathcal{V}_\nu$ such that $F_\phi(D_m) = Q_\phi(\nu(D_m))$ for all datasets $D_m \in \mathcal{D}$.

**Proof.** *Sufficiency.* Suppose that there exists a map $Q_\phi$ such that $F_\phi(D_m) = Q_\phi(\nu(D_m))$ for all $D_m \in \mathcal{D}$. Then, for any two datasets $D_m, D'_m$ satisfying $\nu(D_m) = \nu(D'_m)$, we have $F_\phi(D_m) = Q_\phi(\nu(D_m)) = Q_\phi(\nu(D'_m)) = F_\phi(D'_m)$. Thus $F_\phi$ is objective-invariant. Hence factorization through $\nu(D_m)$ is sufficient for objective-invariant inference.

*Necessity.* Conversely, suppose that $F_\phi$ is objective-invariant. We show that $F_\phi$ must factor through the dataset signature. For any $v \in \mathcal{V}_\nu$, choose any dataset $D_m \in \mathcal{D}$ such that $\nu(D_m) = v$, and

define $Q_\phi(v) := F_\phi(D_m)$. This definition is independent of the choice of $D_m$. Indeed, if $D'_m$ is another dataset satisfying $\nu(D'_m) = v$, then $\nu(D_m) = v = \nu(D'_m)$, and objective-invariance gives $F_\phi(D_m) = F_\phi(D'_m)$. Therefore $Q_\phi$ is well-defined on $\mathcal{V}_\nu$. By construction, $F_\phi(D_m) = Q_\phi(\nu(D_m))$ for every dataset $D_m \in \mathcal{D}$. Hence factorization through $\nu(D_m)$ is necessary for objective-invariant inference.

Combining the two directions, factorization through the objective-induced dataset signature is necessary and sufficient for objective-invariant inference.

**Remark.** The characterization follows directly once the relevant equivalence relation has been identified. The purpose of this characterization is to separate two requirements that are easily conflated. A naive latent-variable learner that ignores visitation decoupling in both the objective and the encoder can entangle transition structure with protocol-induced visitation variation. The VD objective addresses the objective side by inducing a dataset signature and an equivalence relation over datasets, but an amortized encoder is a separate data-dependent map and need not respect that equivalence. Once this objective-induced equivalence relation is identified, the inheritance condition follows directly: the encoder must be constant on its equivalence classes, equivalently, it must factor through the objective-induced dataset signature. Thus Lemma 2 formalizes the objective–inference compatibility requirement under which visitation-decoupled semantics are inherited by amortized inference.

## B Implementation Details and Training Objective

We use PyTorch [18], Gymnasium [26], and MuJoCo-based benchmarks [25] in our numerical experiments. Unless otherwise stated, neural components are optimized with Adam; learning rates, batch sizes, weight decay, training steps, random seeds, and command-line overrides are specified in the supplemental code/configuration files.

Below, this appendix summarizes the concrete training objective and encoder implementations used in our codebase. For each environment $m$, an offline dataset is $\mathcal{D}_m = \{(s_{m,i}, a_{m,i}, s'_{m,i})\}_{i=1}^{N_m}$ with nonnegative weights $w_{m,i} \geq 0$. We write $x_{m,i} := (s_{m,i}, a_{m,i}, s'_{m,i})$ and use a diagonal Gaussian variational posterior.

### B.1 Full-dataset inference and full-batch computation (Toy; also used at evaluation)

**Weight gauge-fixing (mean normalization).** In the implementation, the reconstruction term is self-normalized (a weighted average), so it is invariant to a global rescaling of the weights. However, the KL scaling depends on the total mass, and raw importance weights can carry an arbitrary global scale (e.g., density ratios are only defined up to a multiplicative constant unless explicitly normalized). To remove this ambiguity and improve numerical stability, we apply a simple *gauge-fixing* step:

$$\tilde{w}_{m,i} := \frac{w_{m,i}}{\frac{1}{N_m}\sum_{j=1}^{N_m} w_{m,j} + \varepsilon}, \qquad Z_m := \sum_{i=1}^{N_m} \tilde{w}_{m,i}, \tag{11}$$

with a small $\varepsilon > 0$. This keeps the *relative* weighting unchanged in the self-normalized reconstruction term while tying the KL scaling to an unambiguous mass.

In the Toy experiments of Appendix C, the weights are prescribed on a fixed scale, so we use them directly without applying the per-dataset mean normalization in Eq. (11). Applying this normalization would make the normalized total weight depend on dataset size even for objective-equivalent datasets, undermining the invariance tested in these constructions.

**Weighted DeepSets set encoder.** Let $g_\phi$ be a per-transition embedder (MLP). The code uses a DeepSets-style *weighted mean* pooling,

$$h(\mathcal{D}_m) := \frac{\sum_{i=1}^{N_m} \tilde{w}_{m,i}\, g_\phi(x_{m,i})}{Z_m + \varepsilon}, \tag{12}$$

which is permutation-invariant and well-defined for variable dataset sizes. The variational posterior is parameterized as

$$q_\phi(\theta_m \mid \mathcal{D}_m) = \mathcal{N}\Big(\mu_\phi\big(h(\mathcal{D}_m), Z_m\big),\ \mathrm{diag}\big(\sigma_\phi\big(h(\mathcal{D}_m), Z_m\big)^2\big)\Big). \tag{13}$$

Depending on configuration, the implementation may pass a scalar summary such as $\log Z_m$ to $\mu_\phi$ and/or $\sigma_\phi$.

**Remark.** The objective-aligned dependence condition in Sec. 4.5 holds for any given nonnegative weights. When estimated or stabilized weights are used, the objective and the encoder share the weighted empirical aggregate induced by these weights. The structural condition remains applicable to the resulting objective, while its approximation to the intended reference-distribution objective depends on the weight construction (Sec. 4.7 and Appendix E.3).

**Full-batch VD-ELBO computation.** Let $\ell_{m,i}(\theta_m) := \log p_\psi(s'_{m,i} \mid s_{m,i}, a_{m,i}, \theta_m)$ denote the decoder log-likelihood term (the code optionally predicts $\Delta s$ instead of $s'$). The implemented weighted ELBO contribution is

$$\mathcal{L}_{\mathrm{VD},m}^{(\mathrm{full})} = \frac{\sum_{i=1}^{N_m} \tilde{w}_{m,i}\, \ell_{m,i}(\theta_m)}{Z_m + \varepsilon} - \frac{\beta}{Z_m + \varepsilon}\, \mathrm{KL}(q_\phi(\theta_m \mid \mathcal{D}_m) \,\|\, p(\theta_m))\,. \tag{14}$$

**Evaluation-time inference.** Although continuous-control training uses context subsampling (next subsection), at evaluation we infer the latent belief using the *full offline dataset* available for each protocol by applying (12)–(13) with $\mathcal{D}_m$. This uses all available evidence without changing the learned encoder, and is natural because the encoder is permutation-invariant and supports variable-size inputs.

### B.2 Context subsampling for training (Pendulum / HalfCheetah)

**Training-time surrogate objective via context subsampling.** For continuous-control environments, full-batch training over large offline datasets is computationally expensive. We therefore train using *context subsampling*: at each update step and for each selected environment $m$, we sample a subset (minibatch) $C_m \subset \{1, \dots, N_m\}$ of size $B$. We form a context batch $\mathcal{C}_m := \{x_{m,i}\}_{i \in C_m}$ and the corresponding weights $\{w_{m,i}\}_{i \in C_m}$.

**Context-level gauge-fixing and aggregation.** On a context batch, we apply the same gauge-fixing step within the context:

$$\tilde{w}_{m,i} := \frac{w_{m,i}}{\frac{1}{B}\sum_{j \in C_m} w_{m,j} + \varepsilon}, \qquad \hat{Z}_m := \sum_{i \in C_m} \tilde{w}_{m,i}, \tag{15}$$

and compute the pooled feature

$$\hat{h}(\mathcal{C}_m) := \frac{\sum_{i \in C_m} \tilde{w}_{m,i}\, g_\phi(x_{m,i})}{\hat{Z}_m + \varepsilon}. \tag{16}$$

We then obtain a context-conditioned posterior

$$q_\phi(\theta_m \mid \mathcal{C}_m) = \mathcal{N}\Big(\mu_\phi\big(\hat{h}(\mathcal{C}_m), \hat{Z}_m\big),\ \mathrm{diag}\big(\sigma_\phi\big(\hat{h}(\mathcal{C}_m), \hat{Z}_m\big)^2\big)\Big), \tag{17}$$

using the same encoder architecture as in the full-batch case (with the context statistics substituted).

**Implemented context VD-ELBO.** Given $\theta_m \sim q_\phi(\cdot \mid \mathcal{C}_m)$, we compute a self-normalized reconstruction term and a $1/\hat{Z}_m$-scaled KL term on the context:

$$\widehat{\mathcal{L}}_{\mathrm{VD},m}(\mathcal{C}_m) = \frac{\sum_{i \in C_m} \tilde{w}_{m,i}\, \ell_{m,i}(\theta_m)}{\hat{Z}_m + \varepsilon} - \frac{\beta}{\hat{Z}_m + \varepsilon}\, \mathrm{KL}(q_\phi(\theta_m \mid \mathcal{C}_m) \,\|\, p(\theta_m))\,. \tag{18}$$

We optimize the corresponding *context-subsampled surrogate* objective

$$\max_{\psi,\phi}\ \mathbb{E}_m\, \mathbb{E}_{\mathcal{C}_m \sim \mathcal{S}_B(\mathcal{D}_m)}\Big[\widehat{\mathcal{L}}_{\mathrm{VD},m}(\mathcal{C}_m)\Big], \tag{19}$$

where $\mathcal{S}_B(\mathcal{D}_m)$ denotes the context-subsampling distribution of size $B$. This objective is not claimed to be the exact full-data VD-ELBO; rather, it is a stable stochastic surrogate used for practical training. Crucially, both the surrogate objective and the encoder depend on the context only through the same (stabilized) weighted aggregates (via $\hat{h}$ and $\hat{Z}_m$), preserving objective alignment at the level of the approximation.

**Scaling convention.** We optimize the self-normalized context objective in (18) as a stable stochastic surrogate. This fixes the objective scale at the *per-context* level (weighted averages), which keeps gradient magnitudes comparable across contexts and environments. Any additional constant rescaling of this surrogate primarily affects the effective step size and can typically be compensated by the learning-rate schedule. To ensure fair comparisons, all methods use the same scaling convention and the same context-subsampling protocol.

### B.3 Objective-aligned residual attention set encoder

This subsection describes the details of the objective-aligned residual attention set encoder used in the paper.

**Stage 1 (objective-consistent weighted mean).** Let $h_i := g_\phi(x_i)$ denote per-transition features and let $\tilde{w}_i$ be the stabilized weights used in the objective. Define the (self-normalized) weighted mean

$$m(\mathcal{S}) := \frac{\sum_{i \in \mathcal{S}} \tilde{w}_i\, h_i}{\sum_{j \in \mathcal{S}} \tilde{w}_j + \varepsilon}, \qquad Z(\mathcal{S}) := \sum_{i \in \mathcal{S}} \tilde{w}_i, \tag{20}$$

and write

$$\bar{w}_i := \frac{\tilde{w}_i}{Z(\mathcal{S}) + \varepsilon} \tag{21}$$

for the corresponding normalized weights. Here $\mathcal{S}$ denotes an input set, either a full dataset $\mathcal{D}_m$ at evaluation or a context $\mathcal{C}_m$ during training. This stage matches the dependence structure of the visitation-decoupled objective and recovers the weighted DeepSets encoder when used alone.

**Stage 2 (residual gated refinement without softmax).** To incorporate limited element-wise interaction while preserving objective consistency, we apply a residual gating mechanism based on centered features

$$r_i := \mathrm{LN}\big(h_i - m(\mathcal{S})\big), \tag{22}$$

where LN denotes layer normalization. We compute keys and values as

$$K_i := W_K r_i, \qquad V_i := W_V h_i, \tag{23}$$

and use a learned query vector $q$ to compute element-wise gates

$$e_i := \mathrm{clip}\left(\lambda_s \frac{q^\top K_i}{\sqrt{d}}, -c, c\right), \qquad \alpha_i := \sigma(e_i) \in (0,1), \tag{24}$$

where $\lambda_s = \sigma(\gamma_s)$ is a learned score scale and $c > 0$ is a clipping constant. We then aggregate gated values using the *same* self-normalized weight measure:

$$\tilde{v}(\mathcal{S}) := \frac{\sum_{i\in\mathcal{S}} \bar{w}_i\, \alpha_i\, V_i}{\sum_{i\in\mathcal{S}} \bar{w}_i\, \alpha_i + \varepsilon}. \tag{25}$$

Finally, we residualize this attention-pooled value against the corresponding ungated weighted value mean,

$$v_{\mathrm{base}}(\mathcal{S}) := \sum_{i\in\mathcal{S}} \bar{w}_i\, V_i, \qquad u(\mathcal{S}) := \lambda_r\big(\tilde{v}(\mathcal{S}) - v_{\mathrm{base}}(\mathcal{S})\big), \tag{26}$$

where $\lambda_r = \sigma(\gamma_r)$ is a learned residual scale. Importantly, we do not apply a softmax normalization *across* elements; all pooling remains an expectation, or a gated expectation, under the same stabilized weights as the objective.

**Posterior parameterization.** The final set representation concatenates the global mean and residual refinement,

$$\mathrm{rep}(\mathcal{S}) := [\, m(\mathcal{S}),\, u(\mathcal{S})\,], \tag{27}$$

and the variational posterior is parameterized as

$$q_\phi(\theta \mid \mathcal{S}) = \mathcal{N}\Big(\mu_\phi(\mathrm{rep}(\mathcal{S}), Z(\mathcal{S})),\ \mathrm{diag}(\sigma_\phi(\mathrm{rep}(\mathcal{S}), Z(\mathcal{S}))^2)\Big). \tag{28}$$

This encoder is permutation-invariant by construction and depends on the input set only through weighted aggregates and ratios of weighted aggregates under the same stabilized weights as the learning objective, satisfying objective-aligned dependence at the level of the (full-batch or context-subsampled) surrogate.

### B.4 Downstream planning evaluation details (open-loop CEM)

We evaluate how protocol-dependent latent beliefs affect planning when the learned model is used as a dynamics model. The evaluation is intended as a diagnostic: it does not perform online belief updates, and targets protocol sensitivity (gap) rather than absolute control performance.

**Belief used for planning.** For each dataset (specified by $(\xi, \text{protocol})$), we use the *saved posterior mean* $\mu$ as a fixed latent belief $z$ and do not update it online during planning or execution.

**Planner: open-loop CEM in the learned model.** We plan an open-loop action sequence of horizon $H$ using the Cross-Entropy Method (CEM) under the learned dynamics model conditioned on $z$. At each CEM iteration, we sample a population of action sequences from a factorized Gaussian distribution over time, clamp actions to the environment bounds, roll out the learned model using mean predictions, and update the sampling distribution by refitting to the top-performing elite set.

**Execution in the true environment and protocol gap.** We evaluate the planned open-loop action sequence in the true environment at the same $\xi$. To remove confounds from different initial-state distributions, we evaluate from a fixed initial state for both protocols. For each $(\xi, \text{protocol})$, we repeat planning with multiple CEM random seeds and report the mean return $J(\xi, \text{protocol})$. We use the same set of CEM seeds for both protocols (common random numbers) to reduce evaluation noise. The planning protocol gap reported in the main text is

$$\Delta_{\mathrm{proto}}(\xi) = \left|J(\xi, \mathrm{A}) - J(\xi, \mathrm{B})\right|. \tag{29}$$

Table 1 reports $\Delta_{\mathrm{proto}}(\xi)$ and its average over $\xi$, aggregated across training seeds.

# C Toy Experimental Details

## C.1 Setup

**Dynamics.** We consider a simple BAMDP with linear-Gaussian transition dynamics parameterized by a one-dimensional latent environment variable $\xi \in \mathbb{R}$. States and actions are scalar, $s, a \in \mathbb{R}$, and transitions follow

$$s' = (1 + \alpha\xi)\, s + (\beta\xi)\, a + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{30}$$

where $\alpha$, $\beta$, and $\sigma$ are fixed constants. The latent environment variable $\xi$ governs environment-specific transition dynamics and is treated as unobserved. We use $\xi = 1$, $\alpha = 0.6$, $\beta = 0.8$, and $\sigma = 0.1$.

**Models.** For the decoder, we model $p_\psi(s' \mid s, a, \theta_1)$ as a Gaussian with fixed variance and a mean parameterized by a neural network taking $(s, a, \theta_1)$ as input, with $\theta_1$ denoting the inferred latent parameter. For the encoder, see Appendix B.1.

**Datasets.** We partition the state–action space into a relevant region $\mathcal{R} = [-1, 1]^2$ and an irrelevant region $\mathcal{I} = [3, 5]^2$. Let $\mathcal{T}_\mathcal{R}$ denote 512 transitions sampled from $\mathcal{R}$ and fixed within each run. For each $N \in \{0, 1, 2, 4, \ldots, 4096\}$, we sample $N$ transitions from $\mathcal{I}$ independently of the sets used at other values of $N$, denoted by $\mathcal{I}_N$ with $\mathcal{I}_0 = \emptyset$, and construct

$$\mathcal{D}_1 = \mathcal{T}_\mathcal{R}, \qquad \mathcal{D}'_{1,N} = \mathcal{T}_\mathcal{R} \cup \mathcal{I}_N. \tag{31}$$

Transitions in $\mathcal{R}$ have unit importance weight, whereas transitions in $\mathcal{I}$ have zero importance weight. Thus $\nu(\mathcal{D}_1) = \nu(\mathcal{D}'_{1,N})$ for every $N$, and $\mathcal{D}'_{1,0} = \mathcal{D}_1$.

**Evaluation protocol.** For each encoder architecture and variant, we train $q_\phi$ once on $\mathcal{D}_1$ under the VD-ELBO. After training, we keep $\phi$ fixed and evaluate the same amortized encoder on both $\mathcal{D}_1$ and $\mathcal{D}'_{1,N}$.

**Metric.** We measure posterior sensitivity by the squared Wasserstein-2 distance between diagonal-Gaussian posteriors,

$$W_2^2\big(q_\phi(\theta_1 \mid \mathcal{D}_1),\, q_\phi(\theta_1 \mid \mathcal{D}'_{1,N})\big)\,, \tag{32}$$

as $N$ varies. Means and standard errors are computed over five seeds on the $W_2^2$ scale before log transformation. Values below $10^{-12}$ are displayed at the plotting floor in Fig. 1.

## C.2 Zero-weight construction with a fixed irrelevant-region baseline

We also consider a construction in which the fixed dataset contains objective-zero-weight transitions. Using the same relevant-region transitions $\mathcal{T}_\mathcal{R}$ within each run, we draw 32 irrelevant-region transitions $\mathcal{I}_{32}^{\text{base}}$ and, independently for each $N$, another set $\mathcal{I}_N^{\text{comp}}$ from $\mathcal{I}$. We construct

$$\overline{\mathcal{D}}_1 = \mathcal{T}_\mathcal{R} \cup \mathcal{I}_{32}^{\text{base}}, \qquad \overline{\mathcal{D}}'_{1,N} = \mathcal{T}_\mathcal{R} \cup \mathcal{I}_N^{\text{comp}}. \tag{33}$$

Since all transitions from $\mathcal{I}$ have zero objective weight, $\nu(\overline{\mathcal{D}}_1) = \nu(\overline{\mathcal{D}}'_{1,N})$ for every $N$.

**Remark on the two constructions.** Figure 1 uses $\mathcal{D}_1 = \mathcal{T}_\mathcal{R}$, so the datasets coincide at $N = 0$. In contrast, the construction used in the submitted version, retained in Fig. 4, includes 32 zero-weight transitions in the fixed dataset and independently samples those in the comparison dataset. Thus the datasets differ at $N = 0$ and need not coincide at $N = 32$, and the unconstrained encoder's posterior discrepancy need not increase monotonically for small $N$. Nevertheless, both constructions preserve objective-equivalence and support the same conclusion: unconstrained encoders can respond to objective-invisible transitions, whereas objective-aligned encoders remain invariant up to numerical precision.

For each encoder architecture and variant, we train $q_\phi$ once on $\overline{\mathcal{D}}_1$ and keep it fixed during evaluation. Figure 4 shows the results for $N \in \{32, 64, \ldots, 4096\}$.

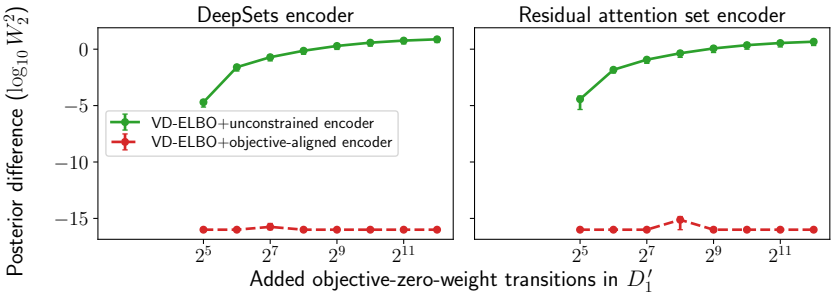


Figure 4: **Toy: zero-weight construction with a fixed irrelevant-region baseline.** The fixed dataset $\overline{\mathcal{D}}_1$ contains 32 zero-weight transitions from $\mathcal{I}$, while $\overline{\mathcal{D}}'_{1,N}$ contains $N$ independently sampled zero-weight transitions from the same region. Here, $N$ denotes the number of zero-weight transitions in the comparison dataset, not the number added to the fixed dataset. The two datasets share the same relevant transitions and are objective-equivalent. We plot $\log_{10}$ of the five-seed mean $W_2^2$ posterior discrepancy for a fixed trained encoder. Unconstrained encoders exhibit posterior discrepancy, whereas objective-aligned encoders remain invariant up to numerical precision. Error bars denote $\pm 1$ standard error across 5 seeds, computed on $W_2^2$ before log transformation.

### C.3 Sanity check with nonzero-weight atom splitting

The zero-weight construction in Appendix C.2 isolates exact objective-equivalence by varying transitions outside the reference support. As an additional sanity check, we consider a nonzero-weight perturbation that also preserves objective-equivalence. For $\eta \in \{0.001, 0.01, 0.1\}$, we assign total importance weight $\eta$ to the 32 irrelevant-region transition atoms in the fixed dataset and $1 - \eta$ to the 512 relevant-region transitions, uniformly within each region. We construct a comparison dataset by replicating the irrelevant-region atoms and splitting their weights so that the total mass assigned to each transition atom is unchanged. Consequently, although the datasets differ as raw transition multisets, including in cardinality and multiplicity, they induce the same weighted empirical measure. Thus, this experiment is not a relaxation of objective-equivalence; it is another objective-equivalence test, now involving nonzero weights in $\mathcal{I}$. The $\eta = 0$ panels instead use the independently sampled zero-weight construction of Appendix C.2.

Under the VD objective, such atom-splitting refinements leave the objective unchanged. An objective-aligned encoder that factors through $\nu(\mathcal{D})$ should therefore remain invariant, whereas an unconstrained encoder with unweighted pooling can still respond to the changed multiplicities or dataset size. Figures 5–6 show that objective-aligned inference remains much less sensitive than unconstrained inference across the tested values of $\eta$.

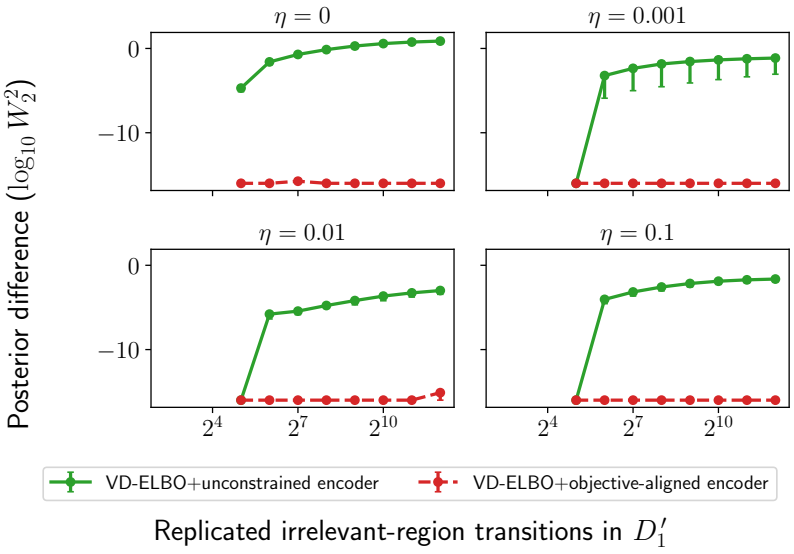


Figure 5: **Toy: sanity check with nonzero-weight atom splitting (deepset).** For $\eta > 0$, the datasets are objective-equivalent by splitting the weights of replicated irrelevant-region transition atoms, preserving their total mass. The $\eta = 0$ panel uses the zero-weight construction of Appendix C.2. We keep the encoder fixed and plot the log-transformed squared Wasserstein-2 posterior distance as the number of irrelevant-region transitions in the comparison dataset increases. The objective-aligned encoder remains much less sensitive than the unconstrained encoder. Error bars denote $\pm 1$ standard error across 5 seeds.

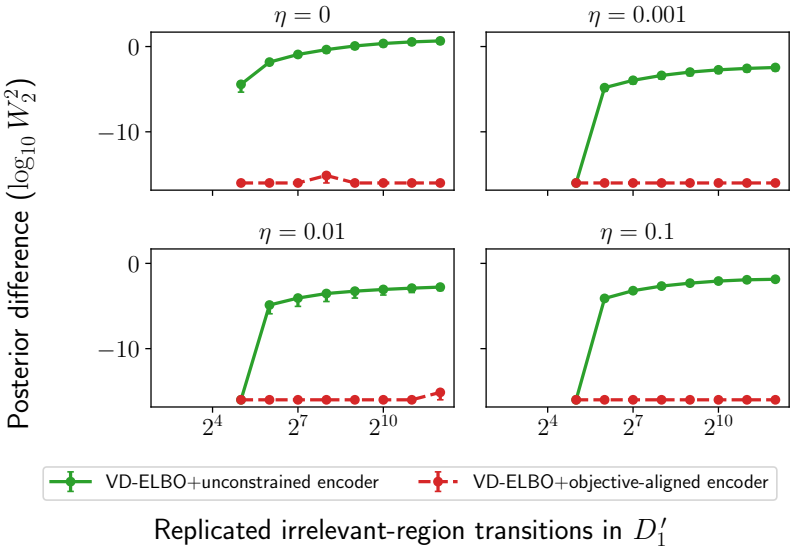


Figure 6: **Toy: sanity check with nonzero-weight atom splitting (residual attention set).** For $\eta > 0$, the datasets are objective-equivalent by splitting the weights of replicated irrelevant-region transition atoms, preserving their total mass. The $\eta = 0$ panel uses the zero-weight construction of Appendix C.2. We keep the encoder fixed and plot the log-transformed squared Wasserstein-2 posterior distance as the number of irrelevant-region transitions in the comparison dataset increases. The objective-aligned encoder remains much less sensitive than the unconstrained encoder. Error bars denote $\pm 1$ standard error across 5 seeds.

# D Pendulum Experimental Details

## D.1 Setup

**Multi-environment setting.** We use PENDULUM-V1 in [26] and vary a single ground-truth dynamics parameter $\xi$ by changing the pendulum length $l$ (i.e., $\xi \equiv l$). We use a three-point grid around the default dynamics, $\xi \in \{0.7, 1.0, 1.3\}$.

**Protocols (offline data collection).** For each fixed $\xi$, we generate two offline datasets using protocols that differ only in the reset distribution. The behavior policy that randomly selects actions, episode horizon, and environment dynamics are otherwise shared. Both protocols are clipped to the same bounded state range:

- **Protocol A:** truncated-Gaussian resets for $(\theta, \dot{\theta})$ with a mean-shifted $\theta$.
- **Protocol B:** uniform resets for $(\theta, \dot{\theta})$ over the default truncated ranges.

We collect $50{,}000$ transitions per dataset. We mainly report the case of data-collecting episode horizon $h = 1$, i.e. a one-step refresh setting where each transition is effectively treated as starting from a reset state. This makes the protocol difference attributable to the reset distribution. We also study the case of $h = 5$ as a $h > 1$ case.

**Reference distribution.** We set the reference distribution $\rho$ to the empirical mixture over all datasets.

**Oracle (analytical) weights.** Since the reset distributions are uniform or factorized truncated normals and $h$=1, the reset-density ratio admits a closed-form expression in log-space (with log-sum-exp stabilization). We use this oracle option purely as a diagnostic to isolate amortization effects from ratio-estimation error.

**Estimated weights (density-ratio estimation).** In the main pendulum setting, we estimate $\log w$ via a binary classifier [23]. Concretely, we train an MLP discriminator to distinguish samples from $\rho$ (label 1) vs. samples from $\mu_m$ (label 0) under a balanced class prior; the discriminator logit then estimates $\log \rho - \log \mu_m$.

**Weight normalization and diagnostics.** Because density ratios are only defined up to a multiplicative constant, we *gauge-fix only the estimated weights* used by our VD-ELBO/objective-aligned encoder by normalizing them to have unit empirical mean within each dataset, i.e., $\tilde{w}_i = w_i / \frac{1}{N} \sum_{j=1}^{N} w_j$ (equivalently, $\sum_i \tilde{w}_i = N$).

**Training and inference.** We train the latent-conditioned dynamics model jointly over all $(\xi, \text{protocol})$ datasets using the context-subsampled VD-ELBO surrogate described above. At evaluation time, we infer each dataset posterior from the full offline dataset and plot the posterior mean against $\xi$ separately for the two protocols. The diagnostic checks whether posterior means track the physical knob $\xi$ while remaining aligned across protocols at fixed $\xi$.

**Reward for planning.** We use the standard PENDULUM-V1 reward computed from the observation $(\cos\theta, \sin\theta, \dot{\theta})$ and action $u$:

$$r = -\left(\theta^2 + 0.1\,\dot{\theta}^2 + 0.001\,u^2\right), \qquad \theta = \operatorname{atan2}(\sin\theta, \cos\theta) \text{ (angle-normalized).} \tag{34}$$

During learned-model rollouts, we renormalize $(\cos\theta, \sin\theta)$ to enforce $\cos^2\theta + \sin^2\theta = 1$.

## D.2 Density-ratio health.

In the pendulum setup, an oracle reset-density ratio $w^\star$ is available, which enables direct validation of the density-ratio estimator. We summarize (i) agreement between estimated log-weights $\log \tilde{w}$ and oracle log-weights $\log w^\star$ ($h = 1$), and (ii) weight-distribution health metrics ($h = 1$ and $h = 5$) in Table 2.

Table 2: **Pendulum density-ratio diagnostics ($h = 1$ vs. $h = 5$).** For $h$=1, Corr/RMSE quantify oracle-vs-estimated agreement in log-space (RMSE is computed after aligning the additive constant). All weight-health metrics (ESS$/N$ and tail statistics) are computed from the *normalized estimated weights* $\tilde{w}$ actually used by our VD-ELBO/objective-aligned encoder, with per-dataset gauge-fixing $\sum_i \tilde{w}_i = N$; $\pm$ denotes the standard deviation over five training seeds (ddof=0), and "worst" denotes the maximum across all seeds.

| Setting | Corr$(\log \tilde{w}, \log w^\star)$ | RMSE$(\log \tilde{w})$ | ESS$/N$ | mean$(\max \tilde{w})$ | worst$(\max \tilde{w})$ |
|---|---|---|---|---|---|
| $h = 1$ | $0.963 \pm 0.003$ | $0.1043 \pm 0.0039$ | $0.784 \pm 0.008$ | $6.927 \pm 0.273$ | 16.2 |
| $h = 5$ | – | – | $0.837 \pm 0.007$ | $3.644 \pm 0.243$ | 6.34 |

### D.3 Posterior Tracking

Below, we report results of posterior tracking for 5 seeds for deepset and residual attention set encoders. Figures 7–8 show results for oracle weights. Figures 9–10 show results for estimated weights with $h = 1$. Figures 11–12 show results for estimated weights with $h = 5$. Each row corresponds to one random seed; the left column uses the unconstrained encoder trained with the standard ELBO, the center column uses the unconstrained encoder trained with the VD-ELBO, and the right column uses the objective-aligned encoder trained with the VD-ELBO. Because latent coordinates are identifiable only up to affine transformations across independent runs, axes are not comparable across rows; interpret *within each row* (within-seed protocol separation/collapse and $\xi$-sensitivity).

The unconstrained encoders trained with the standard ELBO exhibit protocol-dependent offsets, showing that a naive method ignoring visitation-decoupling can actually entangle the two sources of variation: information about transition dynamics and protocol-induced visitation variation. The unconstrained encoders trained with the VD-ELBO remain sensitive to protocol-induced visitation variation, indicating that the objective-level invariance induced by visitation decoupling is not automatically inherited by an unconstrained amortized encoder. In contrast, the objective-aligned encoders largely align the two protocol curves while preserving $\xi$-sensitivity, providing empirical support that the encoder-side inheritance condition is effective in making amortized inference respect the VD objective's invariance.

The three weighting settings serve different diagnostic purposes. The oracle-weight results isolate the amortization effect from weight construction (density-ratio estimation) error: even when the VD objective is implemented with analytical weights, unconstrained encoders can remain protocol-sensitive, whereas objective-aligned encoders inherit the intended invariance. The estimated-weight results with $h = 1$ correspond to the primary pendulum diagnostic analyzed in the main text, where the protocol shift is induced directly by the reset distribution and the weights are estimated from data. The estimated-weight results with $h = 5$ test a less idealized case in which the reset-induced visitation difference propagates through short rollouts. Across these settings, the qualitative pattern is consistent: objective alignment reduces protocol-dependent posterior drift while retaining sensitivity to the environment parameter $\xi$.

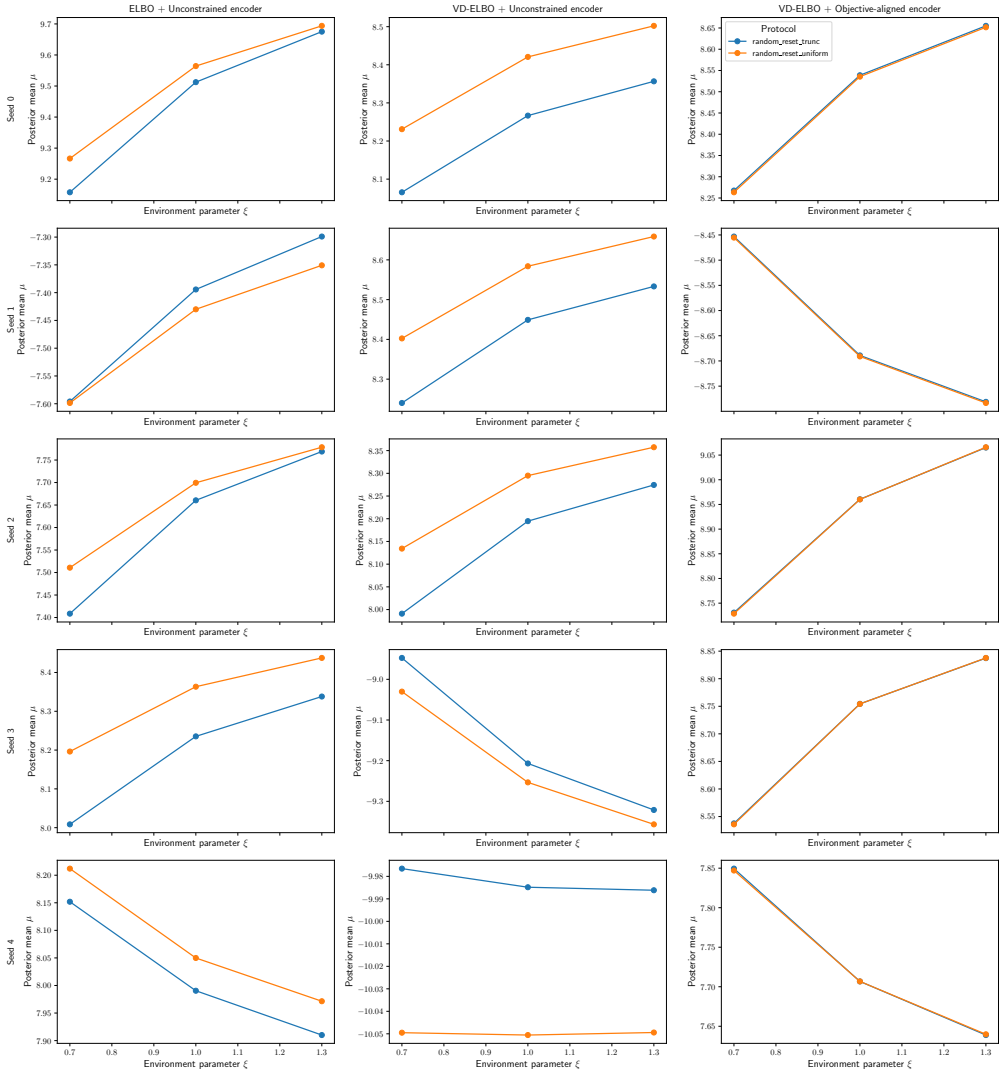


Figure 7: **Pendulum: posterior tracking (deepset, oracle weights, $h = 1$).**

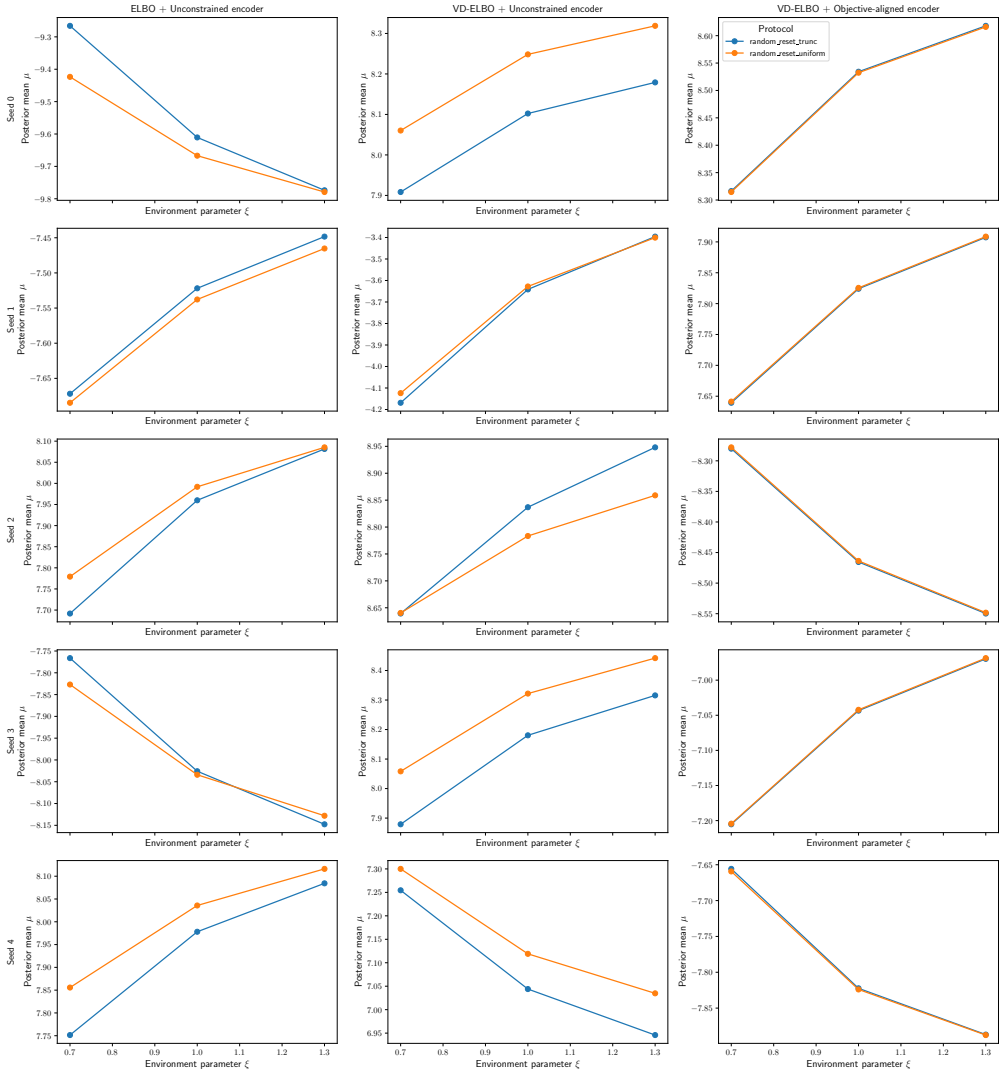


Figure 8: **Pendulum: posterior tracking (residual attention set, oracle weights, $h = 1$).**

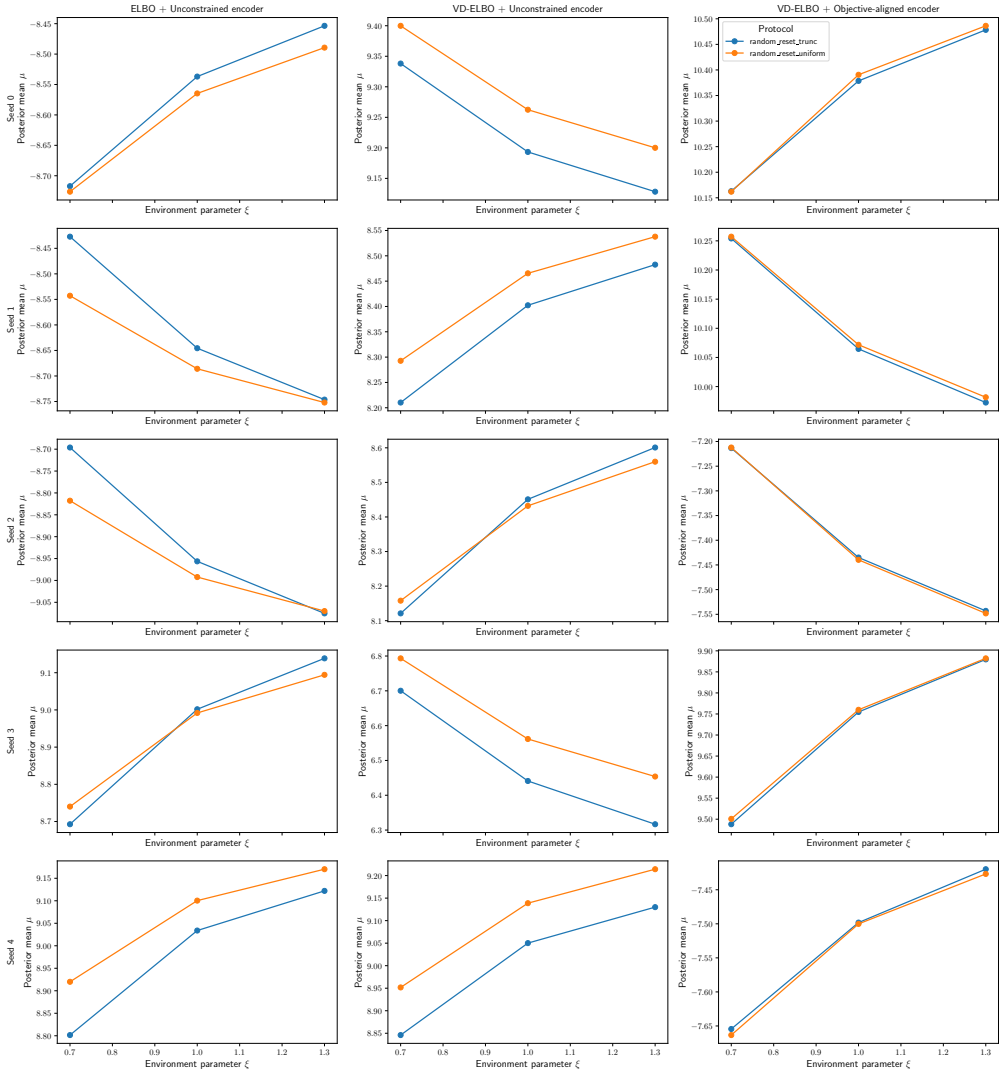


Figure 9: **Pendulum: posterior tracking (deepset, estimated weights, $h = 1$).**

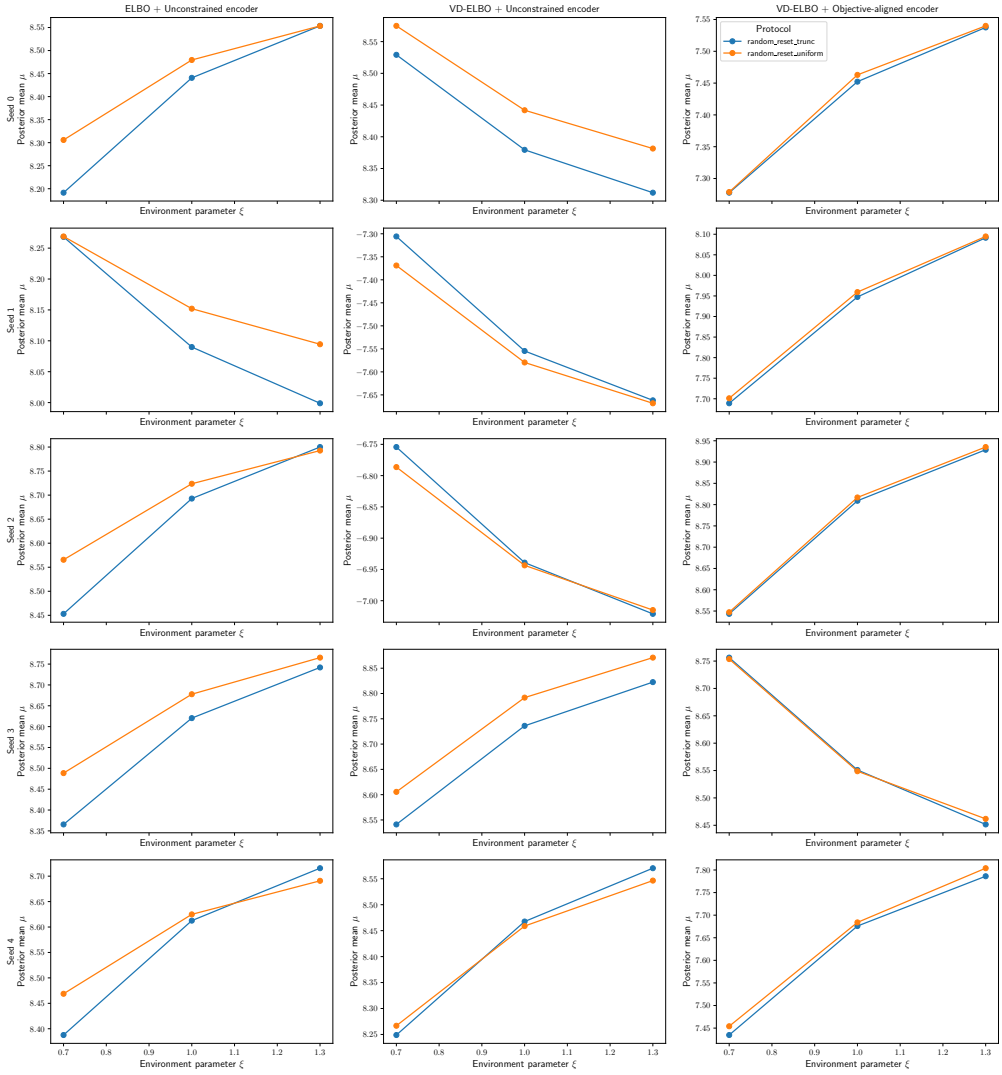


Figure 10: **Pendulum: posterior tracking (residual attention set, estimated weights, $h = 1$).**

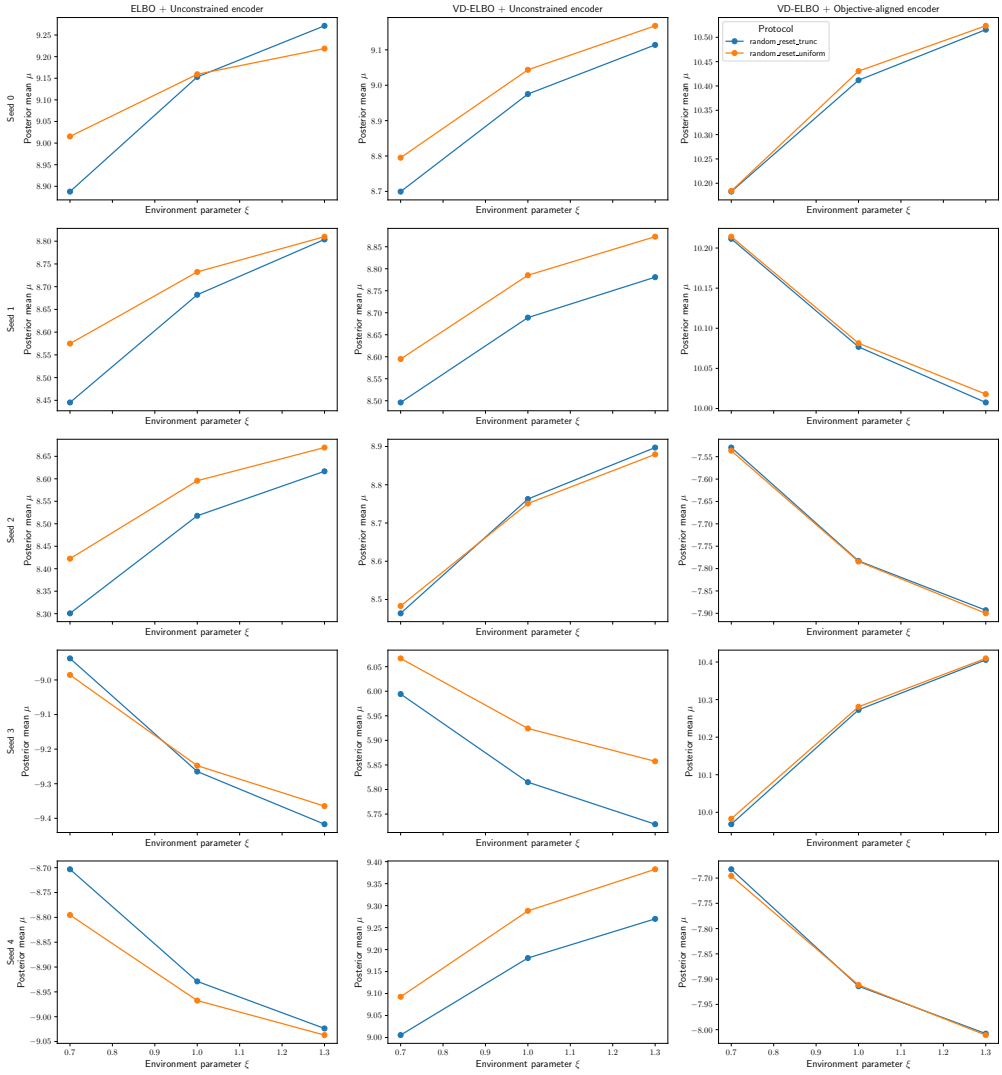


Figure 11: **Pendulum: posterior tracking (deepset, estimated weights, $h = 5$).**

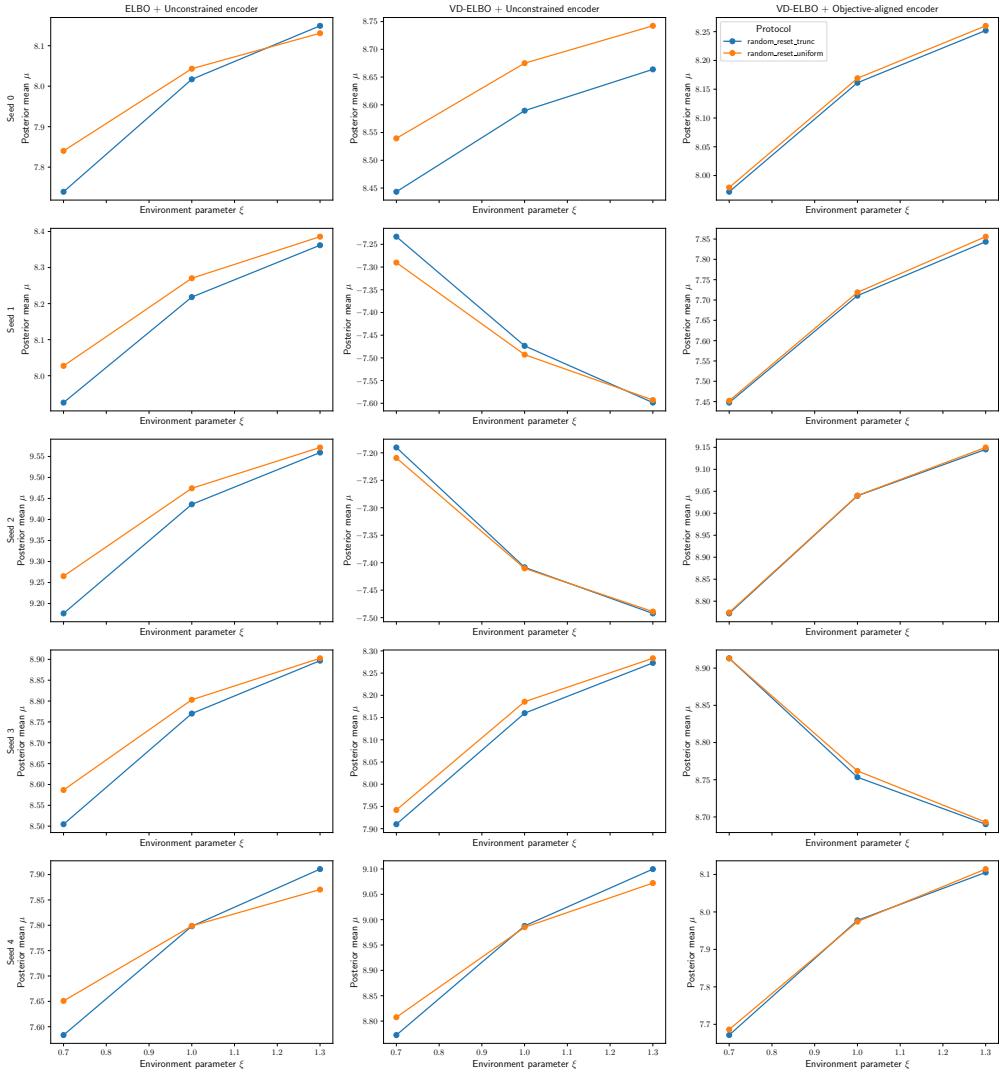


Figure 12: **Pendulum: posterior tracking (residual attention set, estimated weights, $h = 5$).**

Table 3: **Pendulum: scalar summary of posterior tracking (estimated weights, $h = 1$).** Metrics are computed from z-scored posterior means as defined above. Values are mean $\pm$ standard deviation across 5 seeds.

| Encoder | Method | Protocol drift $\Delta_{\text{proto}} \downarrow$ | $\xi$-sensitivity $S_\xi \uparrow$ | Relative drift $R_{\text{proto}} \downarrow$ |
|---|---|---|---|---|
| DeepSets | ELBO + UE | $0.381 \pm 0.170$ | $2.315 \pm 0.058$ | $0.166 \pm 0.078$ |
| DeepSets | VD-ELBO + UE | $0.594 \pm 0.243$ | $2.264 \pm 0.073$ | $0.265 \pm 0.112$ |
| DeepSets | VD-ELBO + OAE | $\mathbf{0.047 \pm 0.014}$ | $2.397 \pm 0.005$ | $\mathbf{0.020 \pm 0.006}$ |
| Residual attention set | ELBO + UE | $0.429 \pm 0.081$ | $2.303 \pm 0.025$ | $0.186 \pm 0.037$ |
| Residual attention set | VD-ELBO + UE | $0.325 \pm 0.230$ | $2.342 \pm 0.054$ | $0.141 \pm 0.103$ |
| Residual attention set | VD-ELBO + OAE | $\mathbf{0.056 \pm 0.027}$ | $2.398 \pm 0.009$ | $\mathbf{0.023 \pm 0.011}$ |

**Scalar summary metrics.** The posterior-tracking plots above are the primary diagnostic, since they show both protocol separation or collapse and sensitivity to the environment parameter within each seed. For completeness, Table 3 reports scalar summaries for the primary estimated-weight setting with $h = 1$. These scalar summaries are intentionally lossy and are meant to support, rather than replace, the visual diagnostics in Figures 9–10.

Let $\mu_p(\xi)$ denote the posterior mean inferred from the full offline dataset collected at environment parameter $\xi$ under protocol $p \in \{A, B\}$, where $A$ and $B$ are the truncated-Gaussian and uniform reset protocols defined in the setup. Because latent coordinates are identifiable only up to affine transformations across independent runs, we first z-score the six posterior means

$$\{\mu_p(\xi) : p \in \{A, B\},\ \xi \in \{0.7, 1.0, 1.3\}\}$$

within each seed and method. Below, $\mu_p(\xi)$ denotes this z-scored posterior mean.

We compute three scalar diagnostics. First, protocol drift measures the average absolute difference between the two protocol posterior means at fixed $\xi$:

$$\Delta_{\text{proto}} = \frac{1}{|\Xi|} \sum_{\xi \in \Xi} |\mu_A(\xi) - \mu_B(\xi)|, \qquad \Xi = \{0.7, 1.0, 1.3\}. \tag{35}$$

Second, $\xi$-sensitivity measures the average within-protocol range of the posterior mean across environment parameters:

$$S_\xi = \frac{1}{2} \sum_{p \in \{A,B\}} \left( \max_{\xi \in \Xi} \mu_p(\xi) - \min_{\xi \in \Xi} \mu_p(\xi) \right). \tag{36}$$

Finally, relative drift normalizes the protocol drift by the amount of $\xi$-sensitivity:

$$R_{\text{proto}} = \frac{\Delta_{\text{proto}}}{S_\xi + \epsilon}, \tag{37}$$

where $\epsilon$ is a small numerical constant. Lower $\Delta_{\text{proto}}$ and $R_{\text{proto}}$ indicate better protocol invariance, while a non-negligible $S_\xi$ checks that the posterior has not collapsed and still tracks the environment parameter.

Table 3 is consistent with the qualitative pattern in the posterior-tracking plots. Across both encoder architectures, objective-aligned inference substantially reduces protocol-dependent posterior drift while preserving sensitivity to the environment parameter. The VD-ELBO with an unconstrained encoder does not consistently remove protocol dependence: it worsens relative drift for DeepSets and improves it only partially for the residual attention set encoder. This supports the claim that visitation decoupling at the objective level does not by itself guarantee inheritance of the induced invariance by amortized inference; the encoder-side objective-alignment condition is needed to obtain consistent protocol invariance.

### D.4 Downstream Planning Discrepancy

In the main text, Table 1 reports protocol discrepancy in downstream planning as the primary diagnostic. This metric compares the achieved return between the two data-collection protocols at

Table 4: **Pendulum: absolute planning returns aggregated over environment parameters (estimated weights, $h$=1; mean $\pm$ std across 5 seeds).** Each entry reports the return averaged over $\xi \in \{0.7, 1.0, 1.3\}$ for a fixed protocol, with variability across training seeds. These returns are reported as an auxiliary sanity check, not as the primary objective-alignment metric. They help verify that the reduced protocol gaps in Table 1 are not explained by a degenerate case in which both protocol-conditioned planners uniformly fail. The absolute returns are broadly comparable across methods; the remaining differences are within the range of outcomes expected from many factors beyond posterior protocol consistency. See Appendix D.4 for discussion.

| Encoder | Method | (trunc-gauss reset) mean return | uniform mean return |
|---|---|---|---|
| DeepSets | ELBO+UE | -368.26 $\pm$ 16.35 | -368.91 $\pm$ 19.43 |
| DeepSets | VD-ELBO+UE | -360.14 $\pm$ 14.58 | -359.96 $\pm$ 16.26 |
| DeepSets | VD-ELBO+OAE | -363.90 $\pm$ 16.90 | -363.87 $\pm$ 16.83 |
| res-attn | ELBO+UE | -360.23 $\pm$ 6.88 | -361.38 $\pm$ 8.05 |
| res-attn | VD-ELBO+UE | -364.12 $\pm$ 8.37 | -364.39 $\pm$ 7.86 |
| res-attn | VD-ELBO+OAE | -360.71 $\pm$ 14.44 | -360.78 $\pm$ 14.40 |

the same environment parameter $\xi$, after planning with the latent belief inferred from each protocol-specific offline dataset. It is therefore aligned with the objective of this paper: testing whether protocol-dependent posterior drift can propagate to downstream model-based control, and whether objective-aligned inference reduces this protocol sensitivity.

For completeness, Table 4 reports the absolute returns obtained by each protocol-specific planner. We include these returns as a sanity check for the protocol-gap diagnostic. The planning gap is the primary metric for objective alignment because it directly measures protocol sensitivity at fixed $\xi$. However, a small gap alone would not imply good planning: two protocol-conditioned planners could achieve similar returns simply because both plans perform poorly. Table 4 rules out this degenerate case in our Pendulum diagnostic. The objective-aligned encoders reduce the protocol gap while keeping absolute returns in the same broad range as the unconstrained baselines, rather than making both protocol-specific planners uniformly fail.

In general, absolute return reflects the full model-based decision-making pipeline: the learned model, the planner or policy, and the true environment all affect the achieved return. Improving absolute return therefore requires these components to work well together. These factors enter absolute return more directly than the protocol gap, because the gap compares two protocol-conditioned planners at the same $\xi$ and can partially cancel common effects. The objective-aligned principle addresses a specific inference-side failure mode under protocol-induced visitation variation, but improving this component alone does not imply that absolute return must improve in every setting. This is consistent with Table 4: objective-aligned encoders substantially reduce the planning discrepancy in Table 1, while their absolute returns remain in a comparable range to the unconstrained baselines, as expected for a targeted inference-side intervention rather than a full end-to-end control optimization.

Based on this distinction, our downstream claim is that reducing protocol-dependent posterior drift can reduce protocol-dependent planning discrepancy without falling into the degenerate uniformly poor-planning case checked above. Accordingly, we report absolute return as a supporting diagnostic, while using the protocol gap as the primary downstream metric for objective alignment.

### D.5 Sensitivity to Reference Mixture Composition

We examine the sensitivity of the Pendulum results to the protocol composition of the reference distribution. Keeping the offline datasets fixed, we set

$$\rho_\lambda = \lambda \rho_{\text{uniform}} + (1-\lambda)\rho_{\text{trunc}}, \qquad \lambda \in \{0.25, 0.50, 0.75\},$$

where each protocol-specific component is the equal mixture over $\xi \in \{0.7, 1.0, 1.3\}$. The case $\lambda = 0.50$ corresponds to the equal-mixture reference used in the main Pendulum experiment. For $\lambda \in \{0.25, 0.75\}$, we re-estimate the importance weights and retrain the VD-ELBO with unconstrained (UE) and objective-aligned (OAE) DeepSets encoders on the same collected datasets. We use the existing Pendulum results for $\lambda = 0.50$. Table 5 reports the posterior-tracking metrics defined above and the protocol-dependent planning gap across five seeds.

Table 5: **Pendulum: sensitivity to the protocol composition of $\rho$ (deepset, estimated weights, $h = 1$).** Values are mean $\pm$ standard deviation across 5 seeds.

| Uniform-protocol mass $\lambda$ | Relative drift UE $\downarrow$ | Relative drift OAE $\downarrow$ | $\xi$-sensitivity OAE $\uparrow$ | Planning gap UE $\downarrow$ | Planning gap OAE $\downarrow$ |
|---|---|---|---|---|---|
| 0.25 | $0.166 \pm 0.110$ | $0.011 \pm 0.003$ | $2.402 \pm 0.006$ | $4.392 \pm 4.828$ | $0.158 \pm 0.170$ |
| 0.50 | $0.265 \pm 0.112$ | $0.020 \pm 0.006$ | $2.397 \pm 0.005$ | $2.160 \pm 1.340$ | $0.320 \pm 0.170$ |
| 0.75 | $0.229 \pm 0.120$ | $0.022 \pm 0.004$ | $2.394 \pm 0.010$ | $2.134 \pm 3.273$ | $0.386 \pm 0.258$ |

Across the tested reference mixtures, OAE consistently reduces relative protocol drift while maintaining $\xi$-sensitivity. OAE also yields a smaller protocol-dependent planning gap than UE in each case. Thus, the improvements in posterior consistency and planning discrepancy are not specific to the equal-mixture reference.

Table 6: **HalfCheetah weight diagnostics ($h$=100).** We report weight-health metrics computed from the *tempered-and-normalized* weights $\tilde{w}$ actually used by our VD-ELBO/objective-aligned encoder, with $\tilde{w}_i \propto w_i^\alpha$ and per-dataset gauge-fixing $\sum_i \tilde{w}_i = N$. "$\pm$" denotes the standard deviation over five training seeds (ddof=0), and "worst" denotes the maximum across all seeds.

| Setting | ESS$/N$ | mean$(\max \tilde{w})$ | worst$(\max \tilde{w})$ |
|---|---|---|---|
| $\alpha = 1.0$ | $0.316 \pm 0.031$ | $161.697 \pm 80.369$ | $1.03 \times 10^3$ |
| $\alpha = 0.9$ | $0.427 \pm 0.035$ | $96.404 \pm 41.648$ | 544 |
| $\alpha = 0.8$ | $0.550 \pm 0.032$ | $57.314 \pm 20.990$ | 281 |
| $\alpha = 0.7$ | $0.672 \pm 0.023$ | $34.088 \pm 10.296$ | 143 |
| $\alpha = 0.6$ | $0.778 \pm 0.013$ | $20.326 \pm 4.906$ | 72.1 |

# E HalfCheetah Experimental Details

## E.1 Setup

**Multi-environment setting.** We use HALFCHEETAH-V5 in [26, 25] and vary a single ground-truth dynamics parameter $\xi$ by scaling the MuJoCo DOF damping terms. We use a three-point grid around the default dynamics, $\xi \in \{0.9, 1.0, 1.1\}$.

**Protocols (offline data collection).** For each fixed $\xi$, we generate two offline datasets using protocols that differ only in the action-sampling distribution. The reset methods, episode horizon, and environment dynamics are otherwise shared. Both protocols are clipped to the same bounded action range to avoid severe support mismatch:

- **Protocol A (Gaussian action policy):** actions are sampled elementwise as $a \sim \mathcal{N}(\mu, \sigma^2)$ with mean $\mu = 0.1$ and standard deviation $\sigma = 0.4$, then clipped to $[-0.8, 0.8]^6$.
- **Protocol B (Uniform action policy):** actions are sampled elementwise as $a \sim \mathrm{Unif}([-0.8, 0.8])$, equivalently $a \sim \mathrm{Unif}([-0.8, 0.8]^6)$.

We collect $50{,}000$ transitions per dataset with episode horizon $h = 100$.

**Reference distribution and weights.** We set the reference distribution $\rho$ to the empirical mixture over all datasets. Importance weights are estimated using the same classifier-based density-ratio procedure as in the pendulum experiment. For the interpolation study, we use tempered weights $(w_{i,m})^\alpha$.

**Training and evaluation.** We train the latent-conditioned dynamics model jointly over all $(\xi, \mathrm{protocol})$ datasets using the context-subsampled VD-ELBO surrogate described above. At evaluation time, we infer each dataset posterior from the full offline dataset and plot the posterior mean against $\xi$ separately for the two protocols. The diagnostic checks whether posterior means track the physical knob $\xi$ while remaining aligned across protocols at fixed $\xi$.

## E.2 Density-ratio health under tempering.

Before inspecting posterior-tracking curves, we summarize the stability of the *tempered-and-normalized* weights used by our objective and objective-aligned encoder. Table 6 reports ESS and tail diagnostics as a function of $\alpha$.

## E.3 Interpreting the interpolation coefficient

In the idealized case where the density ratio $w = d\rho/d\mu$ is known, using the raw weights corresponds directly to the VD semantics specified by the reference distribution $\rho$. Tempering the weights, e.g. using $w^\alpha$ with $\alpha < 1$, changes this effective weighting and can therefore be viewed as moving away from the raw $\rho$-specified semantics. In our HalfCheetah diagnostic, we use tempering in a more limited role: as a simple stabilization family for noisy finite-sample weights.

The density-ratio estimates in HalfCheetah can be noisy or heavy-tailed. We therefore apply the same tempered weights to both the VD-ELBO and the objective-aligned encoder, preserving the compatibility between the objective and the encoder under the chosen stabilized weighting. The sweep over $\alpha$ is intended to test whether approximate objective alignment can remain visible under stabilized noisy weights, not to propose $\alpha = 0.8$ as a universal choice or as a new semantic target.

If tempering is used as a stabilization family, selecting $\alpha$ remains a practical issue. Smaller values of $\alpha$ reduce the tail behavior of the estimated weights and can improve numerical stability, but they also move the effective weighting farther from the raw $\rho$-specified VD semantics. Larger values of $\alpha$ are closer to the raw weighting, but can be more sensitive to density-ratio estimation error, heavy tails, and support mismatch. Thus, choosing $\alpha$ requires a criterion for evaluating the trade-off between weight stability and fidelity to the intended VD semantics. Such criteria may be based on reference predictive performance, balance diagnostics, effective sample size, or tail behavior, but they themselves depend on estimated reference quantities or chosen diagnostic features. Thus, selecting $\alpha$ in a principled way remains a practical open problem, rather than being resolved by a simple validation rule.

More broadly, this is one instance of the weight-construction and stabilization problem discussed in Sec. 6. Tempering is only one simple choice; developing richer and more principled approaches to weight construction, stabilization, and selection is an important direction for future work.

### E.4 Posterior Tracking

Figure 13–17 show deepset results of posterior tracking for 5 seeds. As discussed in Sec. 5.3, the intermediate value, $\alpha = 0.8$, gives relatively stable posterior tracking with reduced protocol-dependent variation. Figure 18–20 observe that similar results are also obtained for the residual attention set encoder.

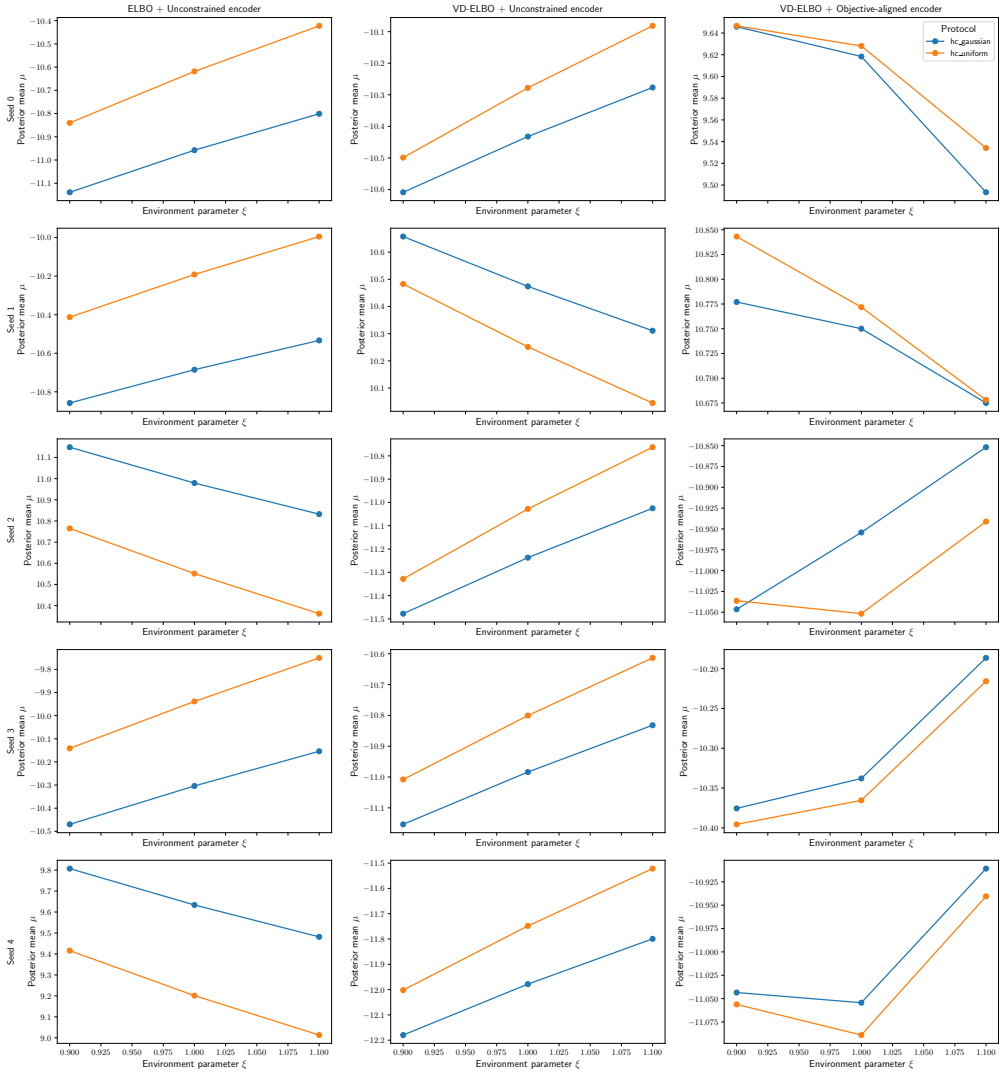


Figure 13: **HalfCheetah: posterior tracking (deepset, $\alpha = 1.0$, estimated weights, $h = 100$).**

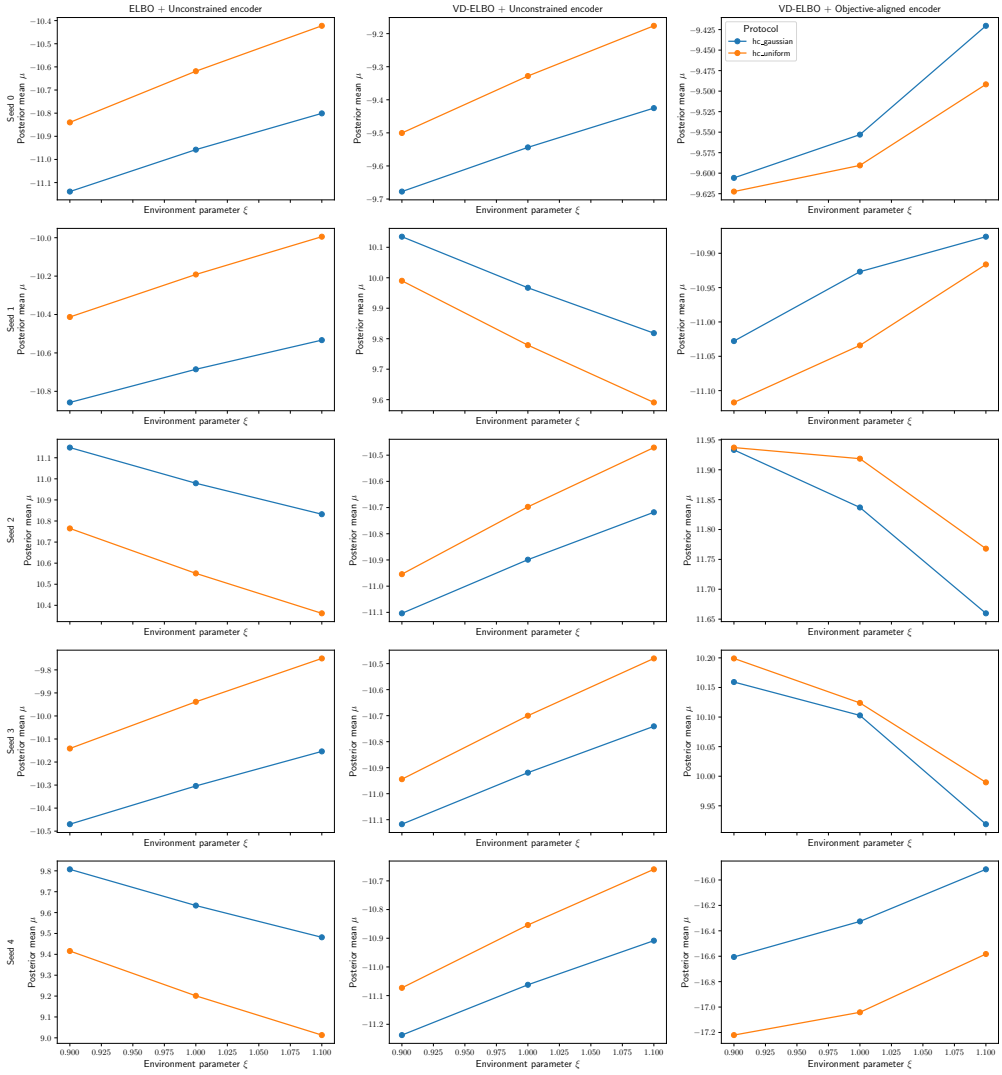


Figure 14: **HalfCheetah: posterior tracking (deepset, $\alpha = 0.9$, estimated weights, $h = 100$).**

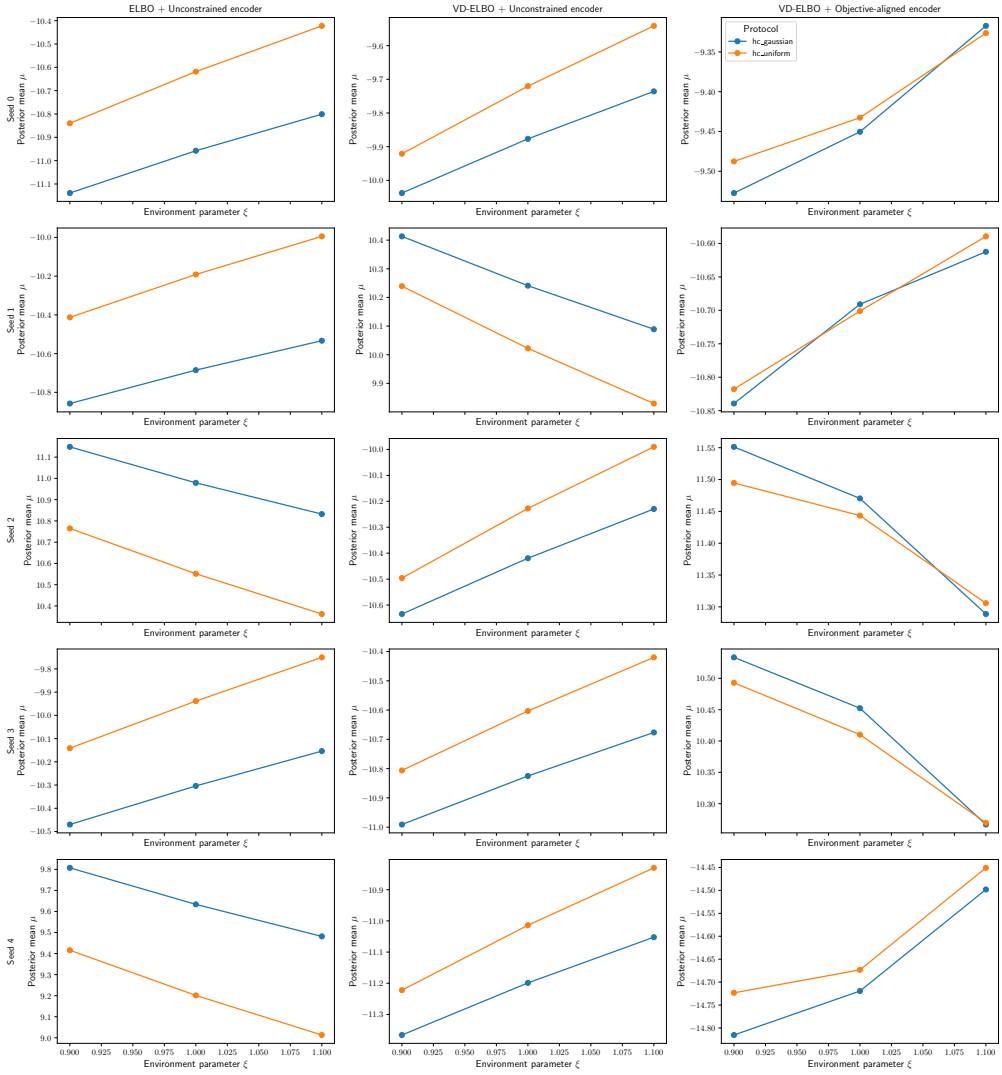


Figure 15: **HalfCheetah: posterior tracking (deepset, $\alpha = 0.8$, estimated weights, $h = 100$).**

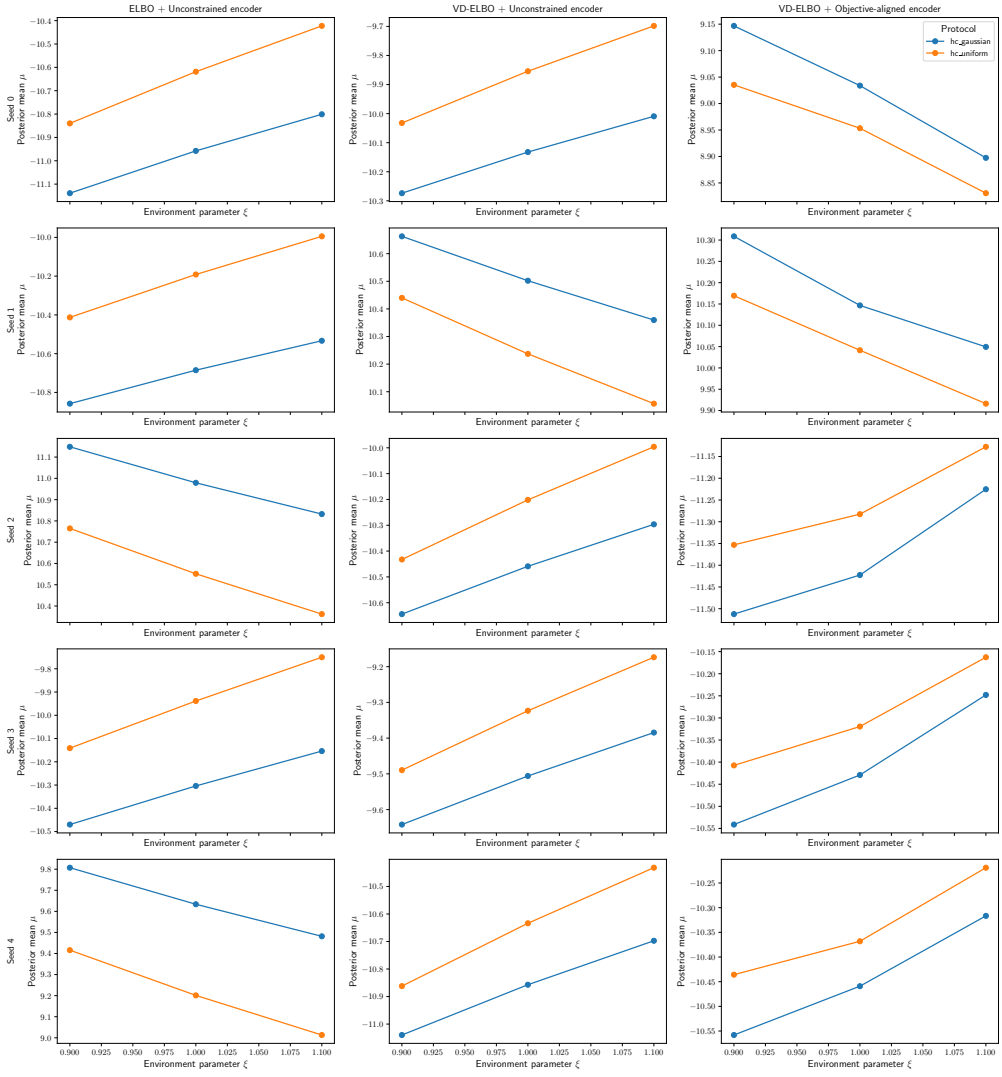


Figure 16: **HalfCheetah: posterior tracking (deepset, $\alpha = 0.7$, estimated weights, $h = 100$).**

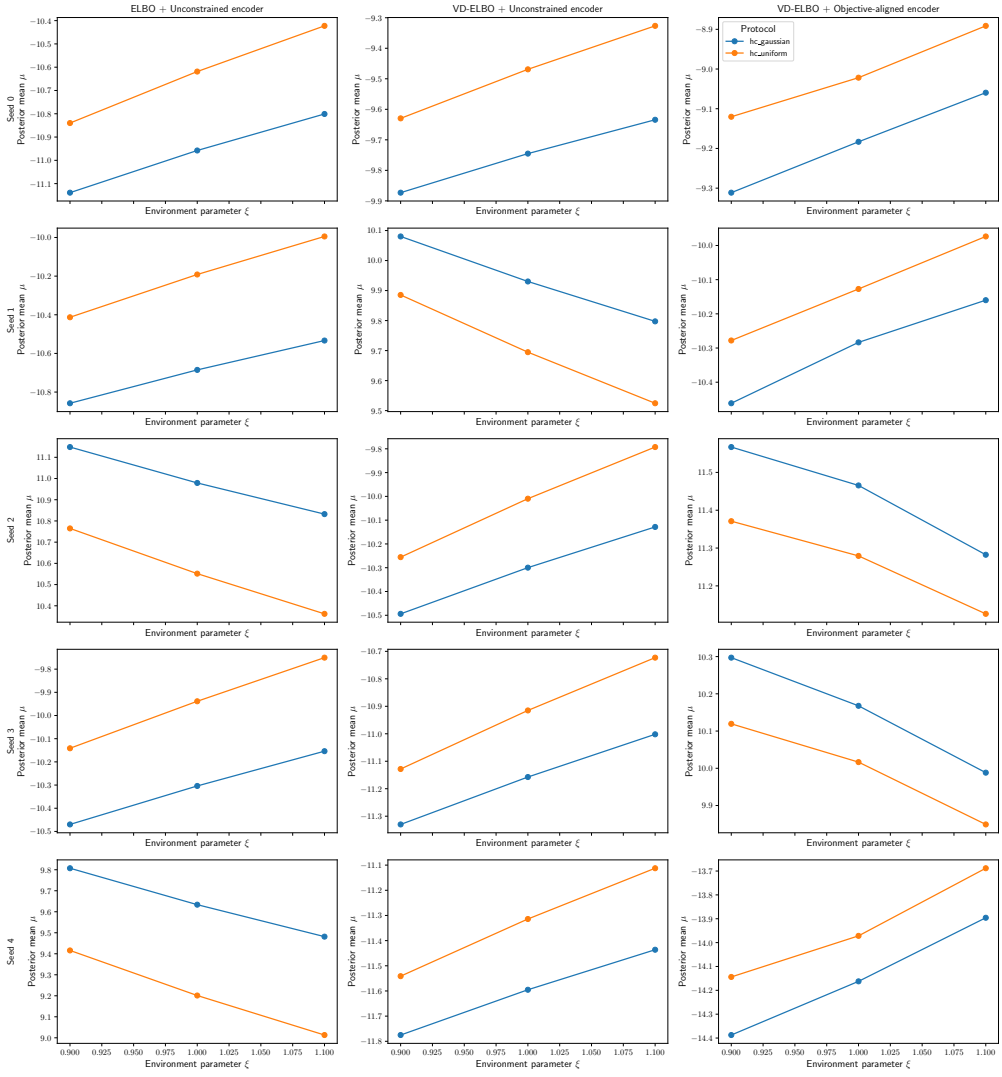


Figure 17: **HalfCheetah: posterior tracking (deepset, $\alpha = 0.6$, estimated weights, $h = 100$).**

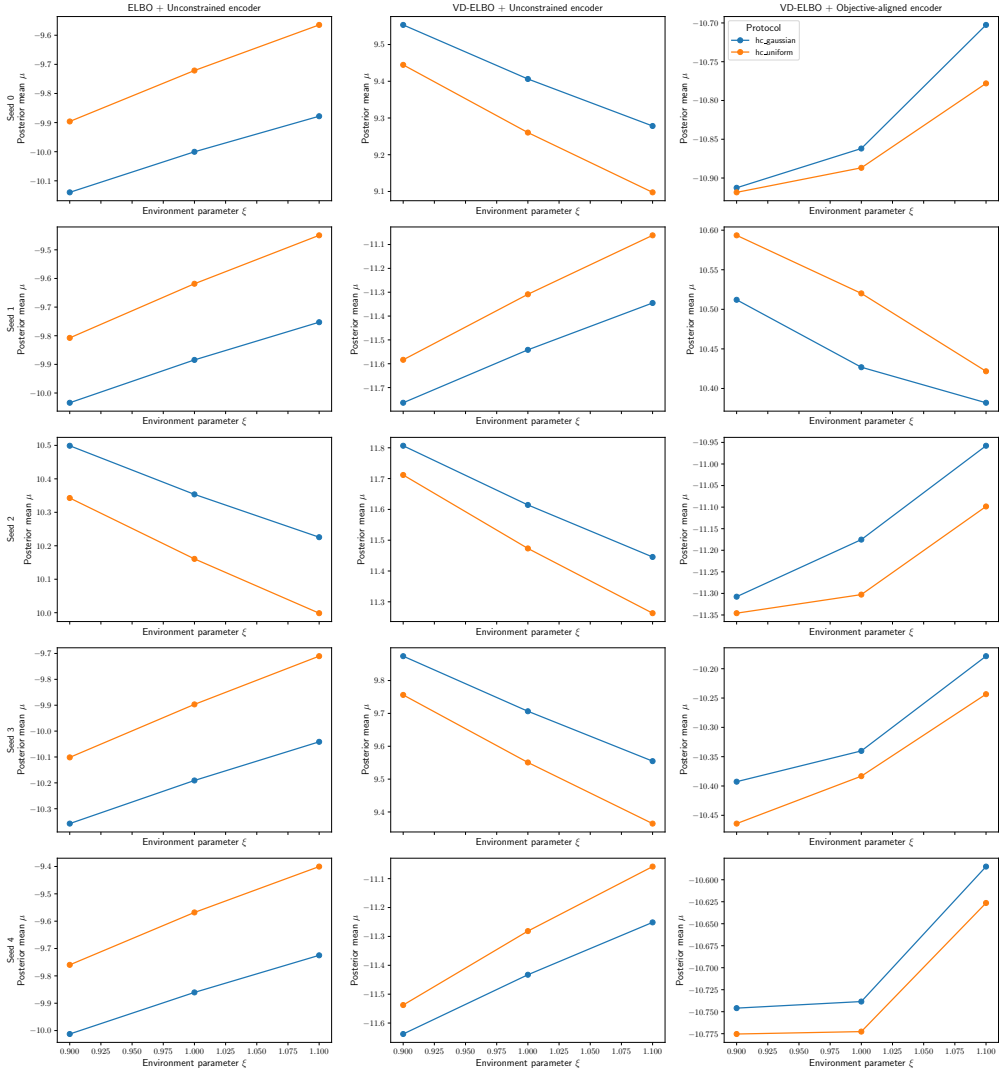


Figure 18: **HalfCheetah: posterior tracking (residual attention set, $\alpha = 0.9$, estimated weights, $h = 100$).**

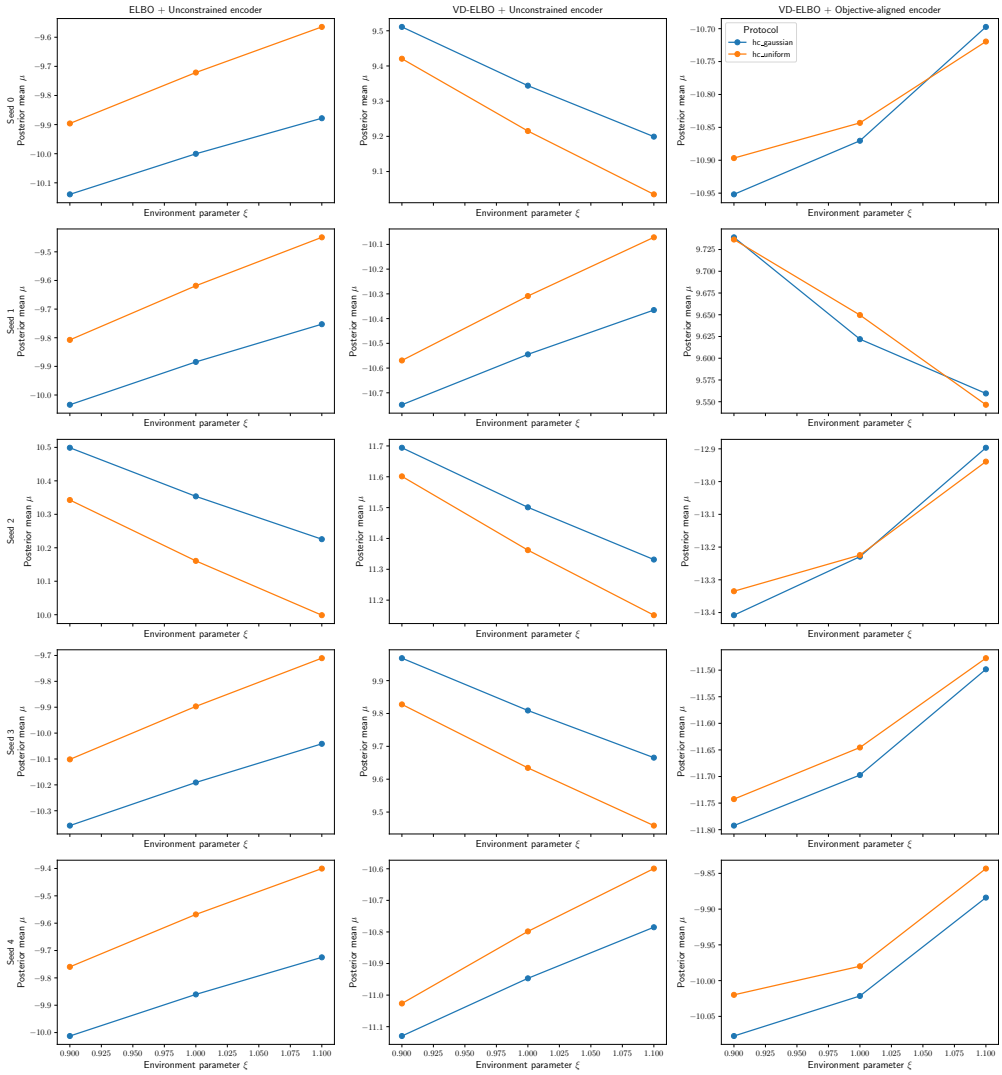


Figure 19: **HalfCheetah: posterior tracking (residual attention set, $\alpha = 0.8$, estimated weights, $h = 100$).**

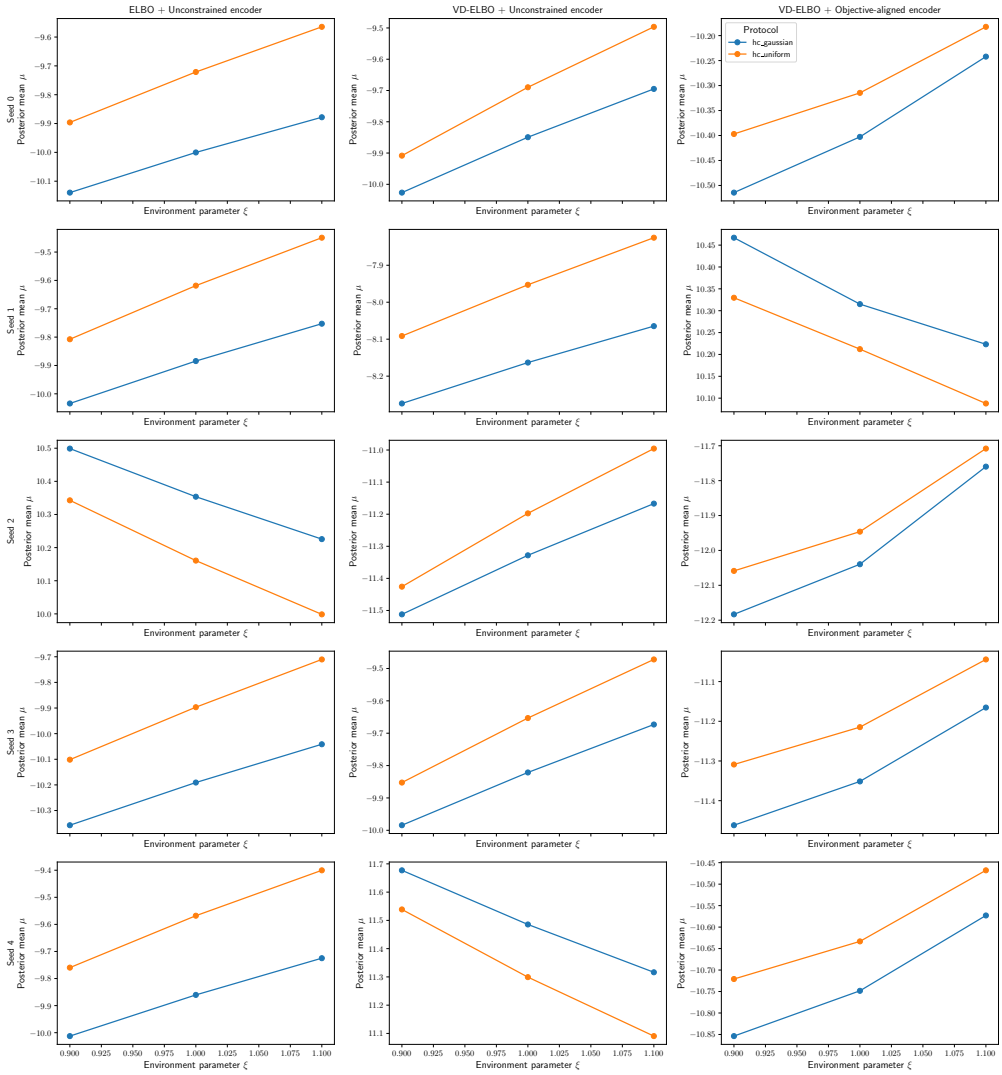


Figure 20: **HalfCheetah: posterior tracking (residual attention set, $\alpha = 0.7$, estimated weights, $h = 100$).**

Table 7: **HalfCheetah: scalar summary of posterior tracking (estimated weights, $\alpha = 0.8$, $h = 100$).** Metrics are computed from z-scored posterior means using the same definitions as in Table 3. Lower protocol drift and relative drift indicate better protocol invariance, while $\xi$-sensitivity checks that the posterior has not collapsed. Values are mean $\pm$ standard deviation across 5 seeds.

| Encoder | Method | Protocol drift $\Delta_{\text{proto}} \downarrow$ | $\xi$-sensitivity $S_\xi \uparrow$ | Relative drift $R_{\text{proto}} \downarrow$ |
|---|---|---|---|---|
| DeepSets | ELBO + UE | $1.605 \pm 0.087$ | $1.442 \pm 0.140$ | $1.126 \pm 0.164$ |
| DeepSets | VD-ELBO + UE | $1.067 \pm 0.131$ | $2.051 \pm 0.100$ | $0.524 \pm 0.089$ |
| DeepSets | VD-ELBO + OAE | $\mathbf{0.316 \pm 0.105}$ | $2.359 \pm 0.056$ | $\mathbf{0.135 \pm 0.048}$ |
| Residual attention set | ELBO + UE | $1.405 \pm 0.117$ | $1.722 \pm 0.131$ | $0.824 \pm 0.122$ |
| Residual attention set | VD-ELBO + UE | $0.914 \pm 0.154$ | $2.156 \pm 0.101$ | $0.428 \pm 0.093$ |
| Residual attention set | VD-ELBO + OAE | $\mathbf{0.339 \pm 0.153}$ | $2.356 \pm 0.062$ | $\mathbf{0.145 \pm 0.070}$ |

**Scalar summary.** For the stabilized interpolation setting $\alpha = 0.8$, we also report scalar summaries of HalfCheetah posterior tracking in Table 7. We use the same posterior-tracking metrics as in Table 3: posterior means are z-scored within each seed and method, protocol drift $\Delta_{\text{proto}}$ measures the average absolute difference between the two protocol posterior means at fixed $\xi$, $\xi$-sensitivity $S_\xi$ measures the within-protocol range across environment parameters, and relative drift $R_{\text{proto}}$ normalizes protocol drift by $\xi$-sensitivity. Here the two protocols correspond to the Gaussian-action and uniform-action data-collection protocols, and $\xi \in \{0.9, 1.0, 1.1\}$ denotes the damping-scale parameter.

Table 7 is consistent with the posterior-tracking plots for $\alpha = 0.8$. Across both encoder architectures, objective-aligned inference substantially reduces protocol-dependent posterior drift while preserving sensitivity to the damping parameter $\xi$. The VD-ELBO with an unconstrained encoder reduces drift relative to the standard ELBO in this setting, but it still leaves substantially larger relative drift than the objective-aligned encoder. This supports the interpretation that, in the noisier HalfCheetah setting, sharing the stabilized weighting between the objective and the encoder helps approximately inherit the intended visitation-decoupled semantics.

Table 8: **HalfCheetah: downstream planning discrepancy ($\alpha = 0.8$, estimated weights, $h$=100; mean $\pm$ std across 5 seeds).**

| Method | $\xi = 0.9$ | $\xi = 1.0$ | $\xi = 1.1$ | mean over $\xi$ |
|---|---|---|---|---|
| deepset: ELBO+UE | $0.6696 \pm 0.3517$ | $1.5448 \pm 1.1972$ | $0.5996 \pm 0.5176$ | $0.9380 \pm 0.5098$ |
| deepset: VD-ELBO+UE | $0.2728 \pm 0.2193$ | $0.5632 \pm 0.2317$ | $0.7405 \pm 0.7523$ | $0.5255 \pm 0.3361$ |
| deepset: VD-ELBO+OAE | $0.3733 \pm 0.1612$ | $0.8662 \pm 1.1284$ | $0.2964 \pm 0.1507$ | $0.5120 \pm 0.3792$ |
| res-attn: ELBO+UE | $0.7463 \pm 0.4633$ | $0.3526 \pm 0.2878$ | $0.4841 \pm 0.3465$ | $0.5277 \pm 0.2684$ |
| res-attn: VD-ELBO+UE | $0.3236 \pm 0.2870$ | $1.2150 \pm 1.2504$ | $0.7123 \pm 0.5669$ | $0.7503 \pm 0.4787$ |
| res-attn: VD-ELBO+OAE | $0.2084 \pm 0.1557$ | $0.2211 \pm 0.2062$ | $0.2049 \pm 0.0325$ | $0.2114 \pm 0.0629$ |

Table 9: **HalfCheetah: absolute planning returns aggregated over environment parameters ($\alpha = 0.8$, estimated weights, $h$=100; mean $\pm$ std across 5 seeds).** Each entry reports the return averaged over $\xi \in \{0.9, 1.0, 1.1\}$ for a fixed protocol, with variability across training seeds. These returns are reported as a supporting diagnostic, not as the primary objective-alignment metric. As in Pendulum, they help interpret the protocol-gap diagnostic by checking whether reduced gaps arise from uniformly poor planning. In this harder HalfCheetah setting, objective-aligned encoders can produce comparable or smaller protocol gaps while attaining lower absolute returns, illustrating that absolute return reflects the full model-based decision-making pipeline rather than posterior protocol consistency alone.

| Encoder | Method | gaussian mean return | uniform mean return |
|---|---|---|---|
| DeepSets | ELBO+UE | $20.138 \pm 1.229$ | $19.365 \pm 1.846$ |
| DeepSets | VD-ELBO+UE | $20.464 \pm 4.226$ | $20.173 \pm 3.847$ |
| DeepSets | VD-ELBO+OAE | $16.905 \pm 4.856$ | $17.108 \pm 5.130$ |
| res-attn | ELBO+UE | $19.869 \pm 2.593$ | $19.642 \pm 2.515$ |
| res-attn | VD-ELBO+UE | $20.015 \pm 3.515$ | $19.448 \pm 2.888$ |
| res-attn | VD-ELBO+OAE | $15.654 \pm 1.735$ | $15.631 \pm 1.774$ |

### E.5 Downstream Planning Discrepancy

Table 8 reports the same downstream planning-discrepancy diagnostic in HalfCheetah. Compared with Pendulum, the planning gaps are smaller and noisier overall, and the differences between methods should be interpreted cautiously. The objective-aligned encoders produce gaps in the same range as the unconstrained baselines, with some reduction in the mean gaps. Thus, the HalfCheetah result is consistent with a mitigation effect, but does not provide the same clean separation as in the Pendulum diagnostic.

This interpretation follows the distinction discussed in Appendix D.4. The protocol gap is less directly affected by the full model-based decision-making pipeline than absolute return, because it compares two protocol-conditioned evaluations at fixed $\xi$ and can partially cancel effects common to both evaluations. In HalfCheetah, however, even this gap diagnostic is harder to evaluate cleanly: the dynamics are higher-dimensional and less smooth, so small action or rollout differences can induce larger trajectory-level changes. Consequently, transition-model error and planner variability can remain visible in the gap diagnostic, making the measured gaps noisier than in Pendulum.

For completeness, Table 9 reports the corresponding absolute returns aggregated over environment parameters. As in Appendix D.4, we report absolute return as a supporting diagnostic while using the protocol gap as the primary downstream metric for objective alignment. The absolute-return results further clarify why the HalfCheetah planning-gap values should not be interpreted as an end-to-end control-performance comparison: objective-aligned encoders can produce protocol gaps in a comparable or smaller range while attaining lower absolute returns. This is consistent with the distinction discussed in Appendix D.4: absolute return reflects the full model-based decision-making pipeline, including the learned model, the planner or policy, and the true environment.

Thus, these results support the weaker interpretation that objective alignment can mitigate downstream planning discrepancy, rather than implying that it necessarily improves planning gaps or absolute returns across all problem settings.

# F Additional GridWorld Experiment

## F.1 Setup

**Dynamics.** We use a stochastic $5 \times 5$ GridWorld with four actions and a latent environment descriptor $\xi \in \{0.05, 0.25, 0.45\}$ controlling the slip probability. The intended action is executed with probability $1 - \xi$, and each of the other three actions with probability $\xi/3$. Actions that would leave the grid keep the agent in the same state. We use a one-dimensional latent variable and a categorical next-state decoder.

**Protocols and weights.** For each $\xi$, we collect two offline datasets of $100{,}000$ transitions with episode horizon $h = 5$. Both protocols use a uniform-random behavior policy but differ in their full-support reset distributions, $0.8\,U(\mathcal{S}_{\text{left/right}}) + 0.2\,U(\mathcal{S})$, where $\mathcal{S}_{\text{left/right}}$ denotes the two leftmost or rightmost columns. We set $\rho$ to the equal mixture of the six exact finite-horizon state–action occupancy distributions $\mu_m$ and compute the ratios $\rho(s, a)/\mu_m(s, a)$ from the tabular transition model. The occupancy ratios are computed exactly; sampled weights are normalized to unit empirical mean within each dataset.

**Training and inference.** Using DeepSets encoders, we compare ELBO with an unconstrained encoder (ELBO+UE), VD-ELBO with an unconstrained encoder (VD-ELBO+UE), and VD-ELBO with an objective-aligned encoder (VD-ELBO+OAE). For each method, we train jointly on the six datasets and infer posteriors using the full datasets at evaluation.

## F.2 Posterior Consistency

We use the same posterior-tracking metrics as in Appendix D, computed from posterior means z-scored within each method and seed over the six datasets. Table 10 reports the mean and standard deviation across five seeds.

Table 10: **GridWorld: posterior-tracking summary (deepset, exact occupancy ratios, $h = 5$).** Mean $\pm$ standard deviation across five seeds.

| Method | Protocol drift $\Delta_{\text{proto}} \downarrow$ | $\xi$-sensitivity $S_\xi \uparrow$ | Relative drift $R_{\text{proto}} \downarrow$ |
|---|---|---|---|
| ELBO + UE | $0.044 \pm 0.026$ | $2.448 \pm 0.001$ | $0.018 \pm 0.011$ |
| VD-ELBO + UE | $0.043 \pm 0.017$ | $2.448 \pm 0.001$ | $0.017 \pm 0.007$ |
| VD-ELBO + OAE | $0.013 \pm 0.002$ | $2.448 \pm 0.001$ | $0.005 \pm 0.001$ |

Across five seeds, objective alignment reduces mean protocol drift by approximately $69\%$ relative to VD-ELBO+UE, with the reduction observed in every seed (Table 11). Meanwhile, $\xi$-sensitivity remains unchanged. These results provide additional evidence that the benefit of objective alignment is not limited to continuous-state environments or Gaussian transition models.

Table 11: **GridWorld: seed-wise protocol drift $\Delta_{\text{proto}}$ (deepset, exact occupancy ratios, $h = 5$).**

| Seed | VD-ELBO + UE | VD-ELBO + OAE |
|---|---|---|
| 0 | 0.02976 | 0.01085 |
| 1 | 0.02420 | 0.01127 |
| 2 | 0.07161 | 0.01638 |
| 3 | 0.04742 | 0.01494 |
| 4 | 0.03974 | 0.01347 |

## G Compute Resources

All experiments were run CPU-only on a single Ubuntu laptop without a compute cluster. The laptop had a 13th Gen Intel(R) Core(TM) i9-13900HX CPU (24 cores, 32 logical CPUs) and 62 GiB RAM. The HalfCheetah experiments required approximately three days of total wall-clock time. The Pendulum and toy experiments together required approximately two days of total wall-clock time.

## H Broader Impacts

This work is foundational research on latent model learning for offline BAMDPs and does not introduce a deployed system, human-subject data, or a high-risk generative model. A potential positive impact is to support more robust model-based decision-making when offline data are collected under heterogeneous protocols. We do not identify direct negative societal impacts specific to the proposed method. As with model-based decision-making methods in general, any downstream deployment should be accompanied by appropriate domain-specific validation and safeguards.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions and scope of the paper, including amortization mismatch and objective-aligned principle. These claims are supported by the theoretical construction in Sec. 4, the empirical evidence in Sec. 5, and the limitations in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Secs. 4.2–4.4 state the assumptions in the counterexample existence lemma. Sec. 5.1 provides the corresponding constructive example. Appendix A.1 gives a formal proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendices B–E describe the experimental protocols and evaluation procedures, and the supplemental code package provides configurations and scripts to reproduce the main toy, Pendulum, and HalfCheetah results. The datasets are generated procedurally, and the reported multi-seed summaries and plots can be regenerated from the released run configurations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplemental material includes anonymized code and instructions for reproducing the main experiments, including dataset generation scripts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendices B–E provide the training objective, encoder implementations, data-collection protocols, weighting procedures, evaluation metrics, and planning details needed to understand the results. Exact hyperparameters and run commands are provided in the supplemental code package.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main quantitative results report seed variability: toy plots use $\pm 1$ standard error across five seeds, and the Pendulum planning-discrepancy table reports mean $\pm$ standard deviation across five seeds. The figure and table captions specify the variability being reported; additional per-seed diagnostics are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report compute resources in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and understood the code of ethics. We do not identify any potential deviations from the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper includes a brief broader-impact discussion in Appendix H.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release such data or models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use PyTorch 2.9.1 (BSD-style license) [18], Gymnasium 1.2.3 (MIT License) [26], and MuJoCo 3.4.0 / MuJoCo-based Gymnasium benchmarks (Apache License 2.0 for MuJoCo; Gymnasium under MIT License) [25]. These assets are properly credited in the paper, and we do not redistribute third-party source code, pretrained models, or external datasets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method does not involve LLMs as important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.