PREFERENCE-GROUNDED LATENT CONTROL: JOINT REPRESENTATION AND POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned latent states underpin world-model agents, embodied robots, and recurrent controllers by determining which information is available for control. Yet conventional latent control optimizes environment prediction or task performance, which need not reflect user preferences. Preference optimization supplies the missing alignment signal, enabling latent states to retain user-relevant distinctions and actors to select preference-consistent actions in otherwise similar situations. Direct Preference Optimization (DPO) is especially attractive because it learns directly from pairwise comparisons without a separate reward model or nested reinforcement-learning procedure. However, standard DPO assumes a shared observable context, whereas latent sequential control must propagate trajectory preferences through an internal state process. Its usual likelihood-ratio reduction therefore cannot jointly train the latent decision state and actor. To address this, we introduce preference sufficiency, which characterizes whether a latent state preserves every distinction required for preference-aligned control, and a representation-error measure for information lost through latent aliasing. We then propose Observation-Anchored Latent Soft-Advantage DPO, a closed-form direct preference framework that anchors the reference policy in observable history space while jointly learning the latent decision state and actor. We prove that the resulting soft-advantage score is canonical and Bellman-consistent, derive an exact actor–representation error decomposition, and show that its representation term vanishes if and only if the state is preference-sufficient. Experiments across sequential-control tasks validate the effectiveness of jointly learning preference-relevant representations and policies.

1 INTRODUCTION

Modern sequential agents increasingly act through learned internal states rather than raw observations or complete histories. Recurrent controllers, representation-based reinforcement learning, and world-model agents compress observation–action histories into latent states used by the actor (Liu et al., 2022; Lenz et al., 2015; Losey et al., 2022; Matsuo et al., 2022; Zhang et al., 2026). These states are typically optimized for reconstruction, dynamics prediction, contrastive learning, or task return (Dutta et al., 2026; Mei et al., 2025; Zheng et al., 2026), rather than for what users prefer. Consequently, current latent control does not ensure that behavior satisfies user preferences. Preference optimization supplies this missing alignment signal, allowing the state to retain user-relevant history and the actor to choose preference-consistent actions. This is fundamentally a representation bottleneck: the actor can exploit only distinctions preserved by the encoder. If histories requiring different preference-optimal actions share the same latent state, no downstream actor can recover the missing information. Preference alignment must therefore shape both the latent decision state and its policy.

Direct Preference Optimization (DPO) is a natural starting point because it turns pairwise comparisons into supervised policy learning through the closed-form solution of KL-regularized reward maximization, without a separate reward model or nested reinforcement-learning procedure (Rafailov et al., 2023). In language models, preferred and rejected responses share a fixed prompt, so the prompt-dependent normalization cancels. Sequential extensions similarly assume supplied states and common environment dynamics, whose transition factors cancel from the policy-to-reference trajectory ratio (Hejna et al., 2024). This reduction does not directly support latent-state learning: preferences label observable rollouts, whereas actions arise through an internal latent process. The resulting ratio contains encoder-dependent inference, transition, and observation factors, while

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

marginalizing latent paths prevents decomposition into per-step action ratios (Lee et al., 2020; Hafner et al., 2025). Representation updates also change the latent coordinate system, making a latent-space reference representation-dependent. A closed-form direct objective must therefore preserve a stable comparison space while preference gradients shape both the state and actor. Together, these challenges raise a fundamental question:

Can trajectory preferences yield a closed-form direct learning principle that jointly aligns latent decision states and policies while the representation itself evolves?

Yes. To this end, we introduce *Observation-Anchored Latent Soft-Advantage DPO* (OA-LSA-DPO), the first closed-form direct preference formulation that jointly learns a history-dependent latent decision state and the actor operating exclusively upon it from trajectory comparisons. Rather than comparing parameter-dependent latent trajectory distributions, we formulate KL-regularized control over observable histories and anchor the reference policy in this encoder-independent space. The current actor remains parameterized by a trainable history encoder, while soft-control optimality yields a closed-form reference-relative latent advantage (Geist et al., 2019; Haarnoja et al., 2017). Accumulating this advantage along each rollout and applying the Bradley–Terry model (Davidson, 1970; Zhang et al., 2024) produces a tractable trajectory-preference objective whose gradients jointly update the encoder and actor. Combined with predictive latent learning, OA-LSA-DPO preserves information for both environment modeling and preference-aligned decisions, without learning a separate reward or value function or evaluating likelihood ratios over learned latent dynamics.

Theoretically, we first identify soft advantage as the unique zero-value canonical representative of a sequential reward-equivalence class, showing that the proposed latent policy ratio is Bellman-consistent rather than a heuristic substitution. We then derive an exact decomposition of reverse-KL policy-matching error into actor approximation and irreducible representation terms, with the latter vanishing exactly under preference sufficiency for the fixed history distribution. Finally, we show that the OA-LSA-DPO objective is invariant under fixed invertible latent reparameterizations, while observation anchoring preserves the encoder preference gradient and avoids the first-order gradient cancellation caused by a co-encoded reference at symmetric initialization. Empirically, we evaluate OA-LSA-DPO against strong direct-preference and reward-model-based baselines under matched realized policy KL. Across diverse sequential-control domains, OA-LSA-DPO consistently improves preference-conditioned control while inducing clearer preference separation in the learned latent space. These results confirm that the improvements arise from jointly adapting the decision representation and actor, rather than from allowing a larger deviation from the reference policy.

Our contributions are summarized as follows:

- We study how to learn both a latent decision state and its policy directly from trajectory preferences. Because the state evolves, standard DPO's fixed-context reduction cannot supervise the encoder. We address this problem with OA-LSA-DPO, the first closed-form direct preference objective for joint latent-state and policy learning. Its observation-space reference remains fixed as the representation evolves, allowing preference gradients to shape the decision state without a separate reward model.
- We establish a theoretical foundation for preference-grounded latent control. We characterize the canonical latent soft advantage, derive an exact actor–representation error decomposition whose representation term vanishes exactly under preference sufficiency, and establish objective invariance under fixed invertible latent reparameterizations and gradient preservation by observation anchoring.
- We evaluate OA-LSA-DPO against frozen-representation direct optimization and reward-model-based preference learning, isolating the effect of preference-driven state construction on both control performance and latent representation quality, thereby demonstrating the advantages of our method.

2 Preliminary

2.1 Latent Sequential Decision Making

We consider a partially observable sequential decision process with environment states $s_t \in \mathcal{S}$, observations $o_t \in \mathcal{O}$, and actions $a_t \in \mathcal{A}$. The environment evolves according to $P(s_{t+1} \mid s_t, a_t)$, and observations are generated from $O(o_t \mid s_t)$. This formulation includes fully observable Markov decision processes as a special case and accommodates settings in which a single observation is insufficient for action selection (Somani et al., 2013; Igl et al., 2018).

At time t, the information available to the agent is the observation–action history $h_t^o := (o_0, a_0, o_1, a_1, \ldots, a_{t-1}, o_t) \in \mathcal{H}$. Rather than conditioning directly on the complete history, latent

sequential agents construct a compact decision state $z_t = f_\eta(h_t^o) \in \mathcal{Z}$. In recurrent implementations, this mapping is realized incrementally as $z_t = F_\eta(z_{t-1}, a_{t-1}, o_t)$, allowing z_t to summarize the information accumulated up to time t. The actor then operates exclusively through this learned representation, with $a_t \sim \pi_\theta(\cdot \mid z_t)$ and composite history policy $\pi_{\eta,\theta}(a_t \mid h_t^o) := \pi_\theta\left(a_t \mid f_\eta(h_t^o)\right)$, where η and θ denote the encoder and actor parameters, respectively. Thus, f_η determines which history-dependent information is available for decision making, whereas π_θ determines how actions are selected from that information. This interface covers recurrent controllers, approximate belief-state policies, representation-based control, and policy-facing latent states in world-model agents (Lenz et al., 2015; Lee et al., 2020; LeCun et al., 2022).

Together with the initial-state distribution and environment dynamics, this composite policy induces a distribution $p_{\eta,\theta}(\tau^o)$ over observable trajectories $\tau^o = (o_0, a_0, o_1, a_1, \ldots, o_{H-1}, a_{H-1}, o_H)$. The corresponding latent sequence $z_{0:H}$ is internal to the agent and is not directly observed by an external preference supervisor.

Latent encoders are commonly trained using reconstruction, dynamics prediction, contrastive learning, or task-return objectives (Gelada et al., 2019; Agarwal et al., 2020; Recanatesi et al., 2021; Teoh et al., 2025). We write these objectives abstractly as

$$\mathcal{L}_{\text{latent}}(\eta, \rho) = \mathbb{E}_{\tau^o \sim \mathcal{D}_{\text{traj}}}\left[\sum_{t=0}^{H-1} \ell_{\text{latent}}\big(z_t, a_t, o_{t+1}; \rho\big)\right], \tag{1}$$

where $z_t = f_\eta(h_t^o)$ and ρ collects any auxiliary dynamics, decoder, or prediction parameters. Depending on the underlying architecture, z_t may be a recurrent hidden state, an approximate belief state, or the policy-facing state of a learned world model (Ha & Schmidhuber, 2018a; Avalos et al., 2024; Wang et al., 2026; Chen et al., 2026). Our framework is agnostic to this choice and requires only that actions be generated through the learned representation, i.e., $\pi_{\eta,\theta}(a_t \mid h_t^o) = \pi_\theta(a_t \mid f_\eta(h_t^o))$.

Crucially, because all action-relevant information must pass through z_t, the encoder determines which distinctions the actor can use. Standard predictive or representation-learning objectives do not specify which distinctions are required for preference-aligned action selection. This motivates allowing trajectory-level preference supervision to shape both the latent decision state and the policy acting through it.

2.2 Direct Preference Optimization and Sequential Feedback

Direct Preference Optimization (DPO) learns a policy from pairwise comparisons without explicitly estimating a reward model (Rafailov et al., 2023). Given a context x, an unknown utility $r^\star(x, y)$, and a fixed reference policy π_{ref}, the solution to KL-regularized reward maximization satisfies

$$r^\star(x, y) = \beta \log \frac{\pi^\star(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log \sum_{y'} \pi_{\text{ref}}(y' \mid x) \exp\left(\frac{r^\star(x, y')}{\beta}\right), \tag{2}$$

where $\beta > 0$ is the regularization coefficient. Under the Bradley–Terry model, preferred and less preferred responses share the same context, so the partition term $\beta \log Z(x)$ cancels from their utility difference (Davidson, 1970; Zhang et al., 2024). This yields the standard DPO objective

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y^w, y^l) \sim \mathcal{D}_{\text{pref}}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y^w \mid x)}{\pi_{\text{ref}}(y^w \mid x)} - \beta \log \frac{\pi_\theta(y^l \mid x)}{\pi_{\text{ref}}(y^l \mid x)}\right)\right]. \tag{3}$$

Sequential feedback instead compares temporally extended behavior. We assume access to pairwise comparisons between observable trajectories, $\mathcal{D}_{\text{pref}} = \{(\tau_i^{o,w}, \tau_i^{o,l})\}_{i=1}^M$, $\tau_i^{o,w} \succ \tau_i^{o,l}$, where $\tau_i^{o,w}$ and $\tau_i^{o,l}$ denote the preferred and less preferred trajectories, respectively. When policies act on a fixed supplied state or history and interact with the same environment, the environment factors cancel from their trajectory-density ratio, leaving a sum of action log-ratios. This cancellation underlies existing direct preference objectives for sequential policies (Hejna et al., 2024); a complete derivation is provided in Appendix B.

The same reduction does not follow by treating a learned latent generative process as the DPO output distribution. Let $p^{\text{lat}}_{\eta,\rho,\theta}(\tau^o)$ denote the observable trajectory likelihood obtained by marginalizing a latent model with representation parameters η, predictive parameters ρ, and actor parameters θ. For current and reference parameters, one generally has

$$\log \frac{p^{\text{lat}}_{\eta,\rho,\theta}(\tau^o)}{p^{\text{lat}}_{\bar{\eta},\bar{\rho},\bar{\theta}}(\tau^o)} \not\equiv \sum_{t=0}^{H-1} \log \frac{\pi_\theta(a_t \mid z_t)}{\pi_{\bar{\theta}}(a_t \mid \bar{z}_t)}, \tag{4}$$

where $z_t = f_\eta(h_t^o)$ and $\bar{z}_t = f_{\bar{\eta}}(h_t^o)$. Latent initialization, transition, and observation factors need not cancel when the representation or predictive model changes. Moreover, preference labels concern observable trajectories rather than particular latent paths, and a ratio of latent-variable marginals does not generally decompose into pointwise action ratios. Different encoders may also assign different coordinates to the same history, making a directly co-encoded reference dependent on the learned representation. Appendix B formalizes these obstacles.

Eq. (4) does not rule out direct preference learning for latent actors. Instead, it shows that the learned latent-model likelihood cannot be substituted directly for the fixed-context likelihood used by standard DPO. The required objective must compare the learned and reference policies on a common, encoder-independent information space while retaining $\nabla_\eta \mathcal{L}_{\text{pref}} \neq 0$. We develop such an observation-anchored construction in Section 3.

3 Preference-Grounded Latent Control

The preceding analysis shows that the difficulty arises from comparing learned and reference policies through parameter-dependent latent trajectory distributions. We address this difficulty by changing the space on which the comparison is defined. Specifically, we formulate the control problem on observable histories, anchor the reference policy in this common information space, and use the learned latent state only to parameterize the current actor. This separation allows environment dynamics to enter through sequential control while preventing changes in the latent representation from altering the meaning of the reference. Building on this principle, we introduce *Observation-Anchored Latent Soft-Advantage DPO* (OA-LSA-DPO), which uses trajectory preferences to jointly shape the latent decision state and the policy acting upon it.

3.1 Observation-Anchored Soft-Advantage Control

A partially observable decision process can be equivalently represented as a fully observed MDP over observation–action histories. Its state space is $\mathcal{H}$, its action space is $\mathcal{A}$, and the environment induces a history transition kernel $P_{\text{env}}(h^{o\prime} \mid h^o, a)$. This representation is theoretical: it does not require the agent to store or process the complete history explicitly. The learned latent state introduced in Section 2.1 will subsequently provide a compact parameterization of a policy on this history space.

We introduce a fixed reference policy $\mu_{\text{ref}}(a \mid h^o)$, defined directly on observable histories. We assume that $\mu_{\text{ref}}(a \mid h^o) > 0$ for every history–action pair appearing in the preference dataset. The reference may be obtained by behavior cloning from the offline trajectory data or by caching the action log-likelihoods of a pretrained history-dependent policy. Crucially, μ_{ref} is fixed during preference optimization and is not parameterized through the learned latent state.

Let $r^\star(h^o, a)$ denote an unknown instantaneous preference utility. This utility is introduced only to characterize the preference-optimal behavior and is never explicitly estimated by our method. Starting from a history h_0^o, consider the KL-regularized sequential control objective

$$J_\beta(\pi; h_0^o) = \mathbb{E}_{\pi, P_{\text{env}}}\left[\sum_{t=0}^{\infty} \gamma^t \left(r^\star(h_t^o, a_t) - \beta \log \frac{\pi(a_t \mid h_t^o)}{\mu_{\text{ref}}(a_t \mid h_t^o)} \right) \middle| \, h_0^o \right], \tag{5}$$

where $\gamma \in [0, 1)$ is the discount factor and $\beta > 0$ controls deviation from the reference policy. Unlike a latent trajectory likelihood ratio, the regularization in Eq. (5) compares action distributions at the same observable history.

Let $V_\beta^\star$ and $Q_\beta^\star$ denote the optimal soft value and soft action-value functions associated with Eq. (5). They satisfy $Q_\beta^\star(h^o, a) = r^\star(h^o, a) + \gamma \mathbb{E}_{h^{o\prime} \sim P_{\text{env}}(\cdot \mid h^o, a)}\left[V_\beta^\star(h^{o\prime})\right]$ and $V_\beta^\star(h^o) = \max_{\pi(\cdot \mid h^o)} \mathbb{E}_{a \sim \pi(\cdot \mid h^o)}\left[Q_\beta^\star(h^o, a) - \beta \log \frac{\pi(a \mid h^o)}{\mu_{\text{ref}}(a \mid h^o)}\right]$. Solving the statewise KL-regularized optimization above yields a closed-form soft value and optimal policy. We provide the complete derivation in Appendix C. In particular, the optimal soft value is

$$V_\beta^\star(h^o) = \beta \log \int_{\mathcal{A}} \mu_{\text{ref}}(a \mid h^o) \exp\left(\frac{Q_\beta^\star(h^o, a)}{\beta}\right) \mathrm{d}a, \tag{6}$$

where the integral is replaced by a sum for discrete action spaces. The optimal policy is

$$\pi_\beta^\star(a \mid h^o) = \mu_{\text{ref}}(a \mid h^o) \exp\left(\frac{Q_\beta^\star(h^o, a) - V_\beta^\star(h^o)}{\beta}\right). \tag{7}$$

Defining the optimal soft advantage as $A^\star_\beta(h^o, a) := Q^\star_\beta(h^o, a) - V^\star_\beta(h^o)$, and rearranging Eq. (7) gives the central reference-relative identity

$$A^\star_\beta(h^o, a) = \beta \log \frac{\pi^\star_\beta(a \mid h^o)}{\mu_{\mathrm{ref}}(a \mid h^o)}. \tag{8}$$

3.2 Preferences as Latent-State Supervision

The observation-anchored identity in Eq. (8) is defined on the complete history space. To obtain a compact decision state, we parameterize the induced history policy through the latent encoder: $\pi_{\eta,\theta}(a \mid h^o) := \pi_\theta\left(a \mid f_\eta(h^o)\right)$. Substituting this composite history policy into the observation-anchored identity gives the latent soft-advantage parameterization

$$A_{\eta,\theta}(h^o, a) := \beta \log \frac{\pi_\theta\left(a \mid f_\eta(h^o)\right)}{\mu_{\mathrm{ref}}(a \mid h^o)}. \tag{9}$$

The numerator is induced by the learned representation and actor, whereas the denominator is evaluated by a fixed reference policy on the original observable history. Thus, both policies are compared on the same history–action event even as the latent representation changes.

Preference gradients shape the decision state. Because the observation-history reference is fixed and independent of the learned encoder, $\nabla_\eta \log \mu_{\mathrm{ref}}(a \mid h^o) = 0$. For a fixed history–action pair, the chain rule therefore gives

$$\nabla_\eta A_{\eta,\theta}(h^o, a) = \beta J_\eta f_\eta(h^o)^\top \left.\nabla_z \log \pi_\theta(a \mid z)\right|_{z=f_\eta(h^o)}, \tag{10}$$

where $J_\eta f_\eta(h^o)$ denotes the Jacobian of the latent state with respect to the encoder parameters. Preference gradients therefore propagate through the actor into the encoder, while the reference remains fixed in an encoder-independent information space. Consequently, optimizing Eq. (9) changes the internal state used by the current policy rather than moving both sides of the comparison together.

Preference-sufficient latent states. We call an encoder f_η *preference-sufficient* for the optimal regularized policy $\pi^\star_\beta$ if there exists a latent policy $\bar{\pi}^\star_\beta$ such that $\pi^\star_\beta(a \mid h^o) = \bar{\pi}^\star_\beta(a \mid f_\eta(h^o))$ for every (h^o, a). Equivalently, $f_\eta(h^o) = f_\eta(\widetilde{h}^o)$ implies $\pi^\star_\beta(\cdot \mid h^o) = \pi^\star_\beta(\cdot \mid \widetilde{h}^o)$. If this condition fails, no actor operating exclusively through $f_\eta(h^o)$ can exactly realize the preference-optimal history policy. Section 4.2 characterizes the resulting irreducible representation error.

Predictive and preference supervision impose complementary requirements. Predictive learning encourages the latent state to retain information required to model the environment, whereas preference supervision encourages it to retain distinctions required for preference-aligned action selection. Preferences therefore do not merely select actions from a learned state; they supervise the construction of the state from which actions are selected. The next subsection aggregates these local latent soft advantages to learn from trajectory-level comparisons.

3.3 Trajectory-Level Direct Preference Optimization

We aggregate the local latent soft advantages in Eq. (9) to learn from trajectory-level comparisons. For an observable rollout τ^o, define

$$S_{\eta,\theta}(\tau^o) := \sum_{t=0}^{H-1} \gamma^t A_{\eta,\theta}(h^o_t, a_t) = \beta \sum_{t=0}^{H-1} \gamma^t \log \frac{\pi_\theta(a_t \mid f_\eta(h^o_t))}{\mu_{\mathrm{ref}}(a_t \mid h^o_t)}. \tag{11}$$

Unlike a likelihood ratio over latent trajectories, this score contains only observation-anchored action ratios; sequential dynamics enter through the soft-control interpretation established in Section 3.1, rather than through learned transition likelihoods. We use the accumulated canonical soft advantage as the trajectory preference score.

Following the Bradley–Terry model introduced in Section 2.2, we assign the comparison probability $P_{\eta,\theta}\left(\tau^{o,w} \succ \tau^{o,l}\right) = \sigma\left(S_{\eta,\theta}(\tau^{o,w}) - S_{\eta,\theta}(\tau^{o,l})\right)$. Maximum likelihood estimation yields our Observation-Anchored Latent Soft-Advantage DPO objective:

$$\mathcal{L}_{\text{OA-LSA-DPO}}(\eta, \theta) = -\mathbb{E}_{(\tau^{o,w}, \tau^{o,l}) \sim \mathcal{D}_{\mathrm{pref}}}\Big[\log \sigma\big(S_{\eta,\theta}(\tau^{o,w}) - S_{\eta,\theta}(\tau^{o,l})\big)\Big]. \tag{12}$$

A single comparison therefore supervises every latent decision state encountered along the two rollouts. The objective requires only observable trajectory labels, propagates sequential credit through both the actor and encoder, and introduces no separate preference reward or value model.

We combine preference-grounded state construction with the predictive latent objective from Section 2.1:

$$\min_{\eta,\rho,\theta} \; \mathcal{L}_{\mathrm{latent}}(\eta, \rho) + \lambda_{\mathrm{pref}} \mathcal{L}_{\text{OA-LSA-DPO}}(\eta, \theta), \tag{13}$$

where $\lambda_{\text{pref}} > 0$. Accordingly, the encoder receives both $\nabla_\eta \mathcal{L}_{\text{latent}}$, which preserves predictive structure, and $\lambda_{\text{pref}} \nabla_\eta \mathcal{L}_{\text{OA-LSA-DPO}}$, which preserves distinctions required for preference-aligned decisions. In short, prediction learns what can happen, whereas preference learns what must be remembered to act.

All preference scores are evaluated on real observable trajectories from $\mathcal{D}_{\text{pref}}$. The learned dynamics contribute only through $\mathcal{L}_{\text{latent}}$; they are neither used to generate imagined preference rollouts nor included in the DPO ratio. Algorithm 1 in Section C summarizes the resulting procedure.

4 Theoretical Analysis

4.1 Sequential Reward Equivalence and Canonical Latent Advantage

We first establish the Bellman foundation of the latent soft-advantage parameterization. Fix the environment transition kernel, temperature $\beta > 0$, and a reference policy $\mu_{\text{ref}}(a \mid h^o)$ with full support. We call two rewards r and $\widetilde{r}$ *control-equivalent*, denoted by $r \simeq_\beta \widetilde{r}$, if they induce the same optimal KL-regularized history policy at every observable history.

Proposition 4.1 (Sequential reward equivalence)**.** *Let $\pi^\star_{\beta,r}$, $V^\star_{\beta,r}$, and $A^\star_{\beta,r}$ denote the optimal regularized policy, soft value, and soft advantage induced by r. For bounded rewards r and $\widetilde{r}$, the following statements are equivalent: (i) $r \simeq_\beta \widetilde{r}$; (ii) $A^\star_{\beta,r}(h^o, a) = A^\star_{\beta,\widetilde{r}}(h^o, a)$ for every (h^o, a); and (iii) there exists a bounded potential $\Psi : \mathcal{H}^o \to \mathbb{R}$ such that*

$$\widetilde{r}(h^o, a) = r(h^o, a) + \gamma \mathbb{E}_{h^{o\prime} \sim P(\cdot \mid h^o, a)} \left[\Psi(h^{o\prime})\right] - \Psi(h^o). \tag{14}$$

Moreover, each control-equivalence class contains a unique representative $r^\circ(h^o, a) := A^\star_{\beta,r}(h^o, a)$ whose optimal soft value is identically zero, $V^\star_{\beta,r^\circ}(h^o) = 0$, and whose optimal soft Q-function therefore satisfies $Q^\star_{\beta,r^\circ}(h^o, a) = r^\circ(h^o, a) = A^\star_{\beta,r}(h^o, a)$.

Proposition 4.1 identifies soft advantage as the unique zero-value canonical reward within each sequential control-equivalence class. When $\gamma = 0$, Eq. (14) reduces to a history-dependent reward offset.

Applying this result to the composite latent policy $\pi_{\eta,\theta}(a \mid h^o) := \pi_\theta(a \mid f_\eta(h^o))$ and its observation-anchored advantage in Eq. (9) gives the following result.

Corollary 4.2 (Bellman consistency of latent log-ratios)**.** *Let $\pi_{\eta,\theta}$ be any normalized composite policy satisfying the required support and integrability conditions. Then the canonical reward $r^\circ_{\eta,\theta}(h^o, a) := A_{\eta,\theta}(h^o, a)$ satisfies $V^\star_{\beta,r^\circ_{\eta,\theta}}(h^o) = 0$, $Q^\star_{\beta,r^\circ_{\eta,\theta}}(h^o, a) = A_{\eta,\theta}(h^o, a)$, and $\pi^\star_{\beta,r^\circ_{\eta,\theta}} = \pi_{\eta,\theta}$.*

Thus, every observation-anchored latent policy ratio defines a valid Bellman-consistent sequential reward. Equation Eq. (9) is therefore not a heuristic substitution, but the canonical soft advantage induced by the complete history-to-latent-to-action policy. Its evaluation requires neither an explicit reward or value model nor likelihood ratios over learned latent transitions. The proofs are provided in Appendix D.

4.2 Exact Decomposition of Latent Preference Error

Preference sufficiency determines whether the optimal history policy can factor through a latent representation, but does not quantify the error introduced by an insufficient encoder. We provide an exact decomposition that separates actor approximation from irreducible representation error.

Let $p_{h^o}(a) := \pi^\star_\beta(a \mid h^o)$ be the preference-optimal policy and let ν be a distribution over observable histories. For an encoder f_η and latent actor $q(\cdot \mid z)$, define $\mathcal{R}_\nu(f_\eta, q) := \beta \mathbb{E}_{h^o \sim \nu} \left[D_{\text{KL}}\left(q(\cdot \mid f_\eta(h^o)) \,\|\, p_{h^o}\right)\right]$. Let ν_η be the distribution of $z = f_\eta(h^o)$ induced by ν. The optimal policies associated with histories mapped to the same z define the normalized geometric barycenter

$$\begin{gathered} m^\nu_\eta(a, z) := \mathbb{E}_\nu[\log p_{h^o}(a) \mid f_\eta(h^o) = z], \quad Z^\nu_\eta(z) := \int_{\mathcal{A}} \exp\left(m^\nu_\eta(a, z)\right) \mathrm{d}a, \\ q^\dagger_{\eta,\nu}(a \mid z) := \frac{\exp\left(m^\nu_\eta(a, z)\right)}{Z^\nu_\eta(z)}. \end{gathered} \tag{15}$$

Theorem 4.3 (Exact latent preference-error decomposition)**.** *Assume that the relevant policies have strictly positive densities with respect to a common base measure and that the conditional*

log-densities in Eq. (15) *are integrable. Then, for every latent actor* q*,*

$$\begin{aligned}\mathcal{R}_\nu(f_\eta, q) =& \mathcal{E}_{\mathrm{act}}(f_\eta, q; \nu) + \mathcal{E}_{\mathrm{rep}}(f_\eta; \nu), \quad \mathcal{E}_{\mathrm{rep}}(f_\eta; \nu) := \beta \mathbb{E}_{z\sim\nu_\eta}\left[-\log Z_\eta^\nu(z)\right], \\ \mathcal{E}_{\mathrm{act}}(f_\eta, q; \nu) &:= \beta \mathbb{E}_{z\sim\nu_\eta}\left[D_{\mathrm{KL}}\left(q(\cdot \mid z) \,\|\, q_{\eta,\nu}^\dagger(\cdot \mid z)\right)\right].\end{aligned} \tag{16}$$

Moreover, $q_{\eta,\nu}^\dagger$ *is the unique risk-minimizing actor* ν_η*-almost surely, and* $\inf_q \mathcal{R}_\nu(f_\eta, q) = \mathcal{E}_{\mathrm{rep}}(f_\eta; \nu)$*. The representation error is zero if and only if* f_η *is preference-sufficient* ν*-almost surely.*

The actor term measures how accurately q realizes the best policy available on the current latent space. For fixed ν, the representation term depends on the encoder and measures disagreement among optimal policies aliased into the same latent state. Actor capacity and optimization can reduce $\mathcal{E}_{\mathrm{act}}$, but not $\mathcal{E}_{\mathrm{rep}}$ while f_η remains fixed. For two equally weighted aliased histories, $q_{\eta,\nu}^\dagger$ reduces to their normalized geometric mean, and the irreducible error is the negative logarithm of their Bhattacharyya coefficient; it is strictly positive unless their optimal action distributions agree almost everywhere.

This decomposition translates directly into control performance. Let $d_\pi^{h_0^o}$ be the normalized discounted history occupancy induced from h_0^o by $\pi(a \mid h^o) = q(a \mid f_\eta(h^o))$.

Corollary 4.4 (Soft-control performance decomposition)**.** *Under the assumptions of Theorem 4.3, for* $\gamma \in [0, 1)$*,*

$$V_\beta^\star(h_0^o) - V_\beta^\pi(h_0^o) = \frac{\mathcal{E}_{\mathrm{act}}(f_\eta, q; d_\pi^{h_0^o}) + \mathcal{E}_{\mathrm{rep}}(f_\eta; d_\pi^{h_0^o})}{1-\gamma} \geq \frac{\mathcal{E}_{\mathrm{rep}}(f_\eta; d_\pi^{h_0^o})}{1-\gamma}. \tag{17}$$

The corollary converts latent aliasing into an on-policy representation penalty over histories visited by π. For a fixed history distribution, actor-only optimization cannot remove this term; its on-policy value may nevertheless change as $d_\pi^{h_0^o}$ changes. Reducing the underlying aliasing requires updating f_η. Complete proofs and the two-history specialization are provided in Appendix E.

4.3 Observation Anchoring under Learnable Representations

Observation anchoring is defined by the independence of the reference policy from the trainable encoder, rather than by a particular latent coordinate system. In particular, the advantage in Eq. (9) uses a fixed history-based reference satisfying $\nabla_\eta \log \mu_{\mathrm{ref}}(a \mid h^o) = 0$. The following proposition formalizes why this property is essential when preferences are used to modify the representation.

Proposition 4.5 (Observation anchoring under latent adaptation)**.** *Assume that* f_η *and* π_θ *are differentiable.*

(i) Objective-level latent-coordinate invariance. *For any fixed diffeomorphism* $\varphi : \mathcal{Z} \to \mathcal{Z}'$*, define* $f_\eta^\varphi := \varphi \circ f_\eta$ *and* $\pi_\theta^\varphi(a \mid z') := \pi_\theta(a \mid \varphi^{-1}(z'))$*. The transformed encoder–actor pair induces the same history policy; hence, its observation-anchored advantage, trajectory score, and OA-LSA-DPO objective are unchanged.*

(ii) Uncancelled encoder gradient. *The encoder gradient of the observation-anchored advantage is given by Eq.* (10)*. It can therefore modify the latent decision state whenever the actor is locally sensitive to that state and the corresponding encoder Jacobian is nonzero.*

(iii) Co-encoded reference cancellation. *Consider instead a reference evaluated through the same trainable encoder,* $A^{\mathrm{co}}_{\eta,\theta,\bar\theta}(h^o, a) := \beta \log \frac{\pi_\theta(a|f_\eta(h^o))}{\pi_{\bar\theta}(a|f_\eta(h^o))}$*. Its encoder gradient and its value at reference initialization satisfy*

$$\begin{aligned}\nabla_\eta A^{\mathrm{co}}_{\eta,\theta,\bar\theta}(h^o, a) &= \beta J_\eta f_\eta(h^o)^\top \left[\nabla_z \log \pi_\theta(a \mid z) - \nabla_z \log \pi_{\bar\theta}(a \mid z)\right]_{z=f_\eta(h^o)}, \\ \theta = \bar\theta &\Longrightarrow A^{\mathrm{co}}_{\eta,\bar\theta,\bar\theta}(h^o, a) = \nabla_\eta A^{\mathrm{co}}_{\eta,\bar\theta,\bar\theta}(h^o, a) = 0.\end{aligned} \tag{18}$$

Thus, a reference that shares the trainable encoder produces no first-order encoder supervision at symmetric initialization. Although subsequent actor updates may break this symmetry, the initial preference signal cannot directly separate aliased histories. In contrast, an observation-history reference, or a reference implemented through a separately frozen encoder, remains independent of η and avoids this structural cancellation.

Corollary 4.6 (Preference separation under transient-cue aliasing)**.** *Consider two histories* $h^{o,0}$ *and* $h^{o,1}$ *that differ only in a binary preference cue* $c \in \{0, 1\}$ *revealed once earlier in the episode. Suppose that they have identical current physical observations and future physical transition laws, while the final decision reward is* $r_c(a) = \mathbf{1}\{a = c\}$ *for* $a \in \{0, 1\}$*, with a uniform reference policy. Let* ν *assign equal probability to the two histories. A collapsed representation* f_{col} *that discards* c

Table 1: **Observation anchoring prevents first-order encoder-gradient cancellation.** Mean ± SEM over five seeds at symmetric initialization.

Reference construction	Encoder-gradient decomposition			Cancellation diagnostic	
	Policy branch $\|\nabla_\eta\|_2$	Reference branch $\|\nabla_\eta\|_2$	Total encoder gradient ↑	Branch cosine	Cancellation fraction ↓
Co-encoded latent reference	6.454± 0.128	6.454± 0.128	0.000	−1.000	1.000
Observation-history anchor (ours)	6.454± 0.128	**0.000**	**6.454± 0.128**	—	**0.000**

Table 2: **Preference-grounded control in MiniGrid.** Mean ± SEM over five seeds.

Method	Preference-conditioned behavior				Decision
	Target success ↑	Balanced ↑	Correct reversal ↑	Pref. accuracy ↑	Junction $p(a^\star)$ ↑
Frozen-latent DPO	0.500± 0.000	0.000± 0.000	0.000± 0.000	0.500± 0.000	0.485± 0.005
Frozen-latent RLHF	0.500± 0.000	0.117± 0.117	0.000± 0.000	0.500± 0.000	0.491± 0.008
Adaptive-latent RLHF	0.667± 0.167	0.333± 0.333	0.333± 0.333	0.667± 0.167	0.599± 0.113
OA-LSA-DPO	**1.000± 0.000**	**1.000± 0.000**	**1.000± 0.000**	**0.993± 0.007**	**0.828± 0.031**

Table 3: **Preference-grounded control in ManiSkill.** Mean ± SEM over five seeds.

Method	Preference-conditioned physical control				Preference representation	
	Target ↑	Balanced ↑	Wrong target ↓	Reversal ↑	Pref. acc. ↑	Preference-subspace separation ↑
Dyn-only LSA-DPO	0.196± 0.105	0.075± 0.052	0.208± 0.112	0.000± 0.000	0.643± 0.013	0.000± 0.000
Frozen-latent RLHF	0.342± 0.048	0.150± 0.076	0.342± 0.054	0.000± 0.000	0.643± 0.015	0.000± 0.000
Adaptive-latent RLHF	0.462± 0.050	0.242± 0.121	0.425± 0.029	0.033± 0.017	0.831± 0.017	2.767± 0.479
OA-LSA-DPO	**0.962± 0.034**	**0.922± 0.053**	**0.000± 0.000**	**1.000± 0.000**	**0.865± 0.014**	**9.665± 1.526**

and the one-bit representation $f_{\mathrm{pref}}(h^{o,c}) = c$ *satisfy* $\inf_q \mathcal{R}_\nu(f_{\mathrm{col}}, q) = \beta \log\cosh\left(\frac{1}{2\beta}\right) > 0$ *and* $\inf_q \mathcal{R}_\nu(f_{\mathrm{pref}}, q) = 0$.

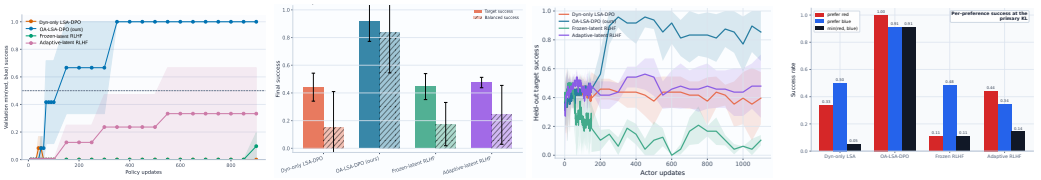
The corollary shows that predictive equivalence does not imply preference equivalence: histories with identical physical futures can still require different actions. Observation anchoring preserves the gradient path needed to separate such histories. Proofs are provided in Appendix F.

5 Experiments

We evaluate whether trajectory preferences can reshape a predictive latent state into a preference-sufficient decision state, and whether observation anchoring is necessary for this representation learning. We use **MiniGrid** as a controlled test of history aliasing with a transient preference cue, and **ManiSkill** as a continuous two-cube manipulation benchmark with heterogeneous offline preferences. Additional details are provided in Appendix G. We compare four preference-optimization methods. *Dyn-only LSA-DPO* freezes the pretrained predictive representation and updates only the policy using a trajectory-level DPO-style preference objective (Hejna et al., 2024). OA-LSA-DPO jointly adapts the preference-facing latent state and policy through our observation-anchored latent soft-advantage objective, without learning an explicit reward model. *Frozen-latent RLHF* keeps the representation fixed, fits a Bradley–Terry reward model from trajectory preferences, and optimizes the policy against the learned reward (Christiano et al., 2017; Ouyang et al., 2022). *Adaptive-latent RLHF* uses the same reward-model-based pipeline while also adapting the preference-facing latent representation.

Observation Anchoring Preserves Representation Supervision. Table 1 verifies Proposition 4.5: a co-encoded reference produces equal-and-opposite encoder-gradient branches, whereas observation anchoring preserves the full policy-side preference gradient at symmetric initialization. When both sides share the trainable encoder, the preference comparison moves with the representation and erases the signal needed to separate aliased histories. Anchoring the reference in observable history removes this moving target while leaving the current latent policy trainable, thereby enabling direct preference supervision of the decision state.

Preference-Grounded Control under History Aliasing. In MiniGrid, identical later observations require opposite actions under different transient requests. Table 2 shows that frozen representations reuse the same decision and Adaptive RLHF recovers this dependence only partially: they may complete the task without reliably remembering which goal was requested. OA-LSA-DPO instead produces correct, balanced, and counterfactually reversible behavior because preference gradients



(a) MiniGrid learning dynamics **(b)** ManiSkill final success **(c)** ManiSkill learning dynamics **(d)** ManiSkill per-profile success

Figure 1: **Preference-conditioned control and learning dynamics.** OA-LSA-DPO achieves high balanced performance in both environments and learns to respond to both preference profiles, whereas frozen and reward-model baselines remain substantially less reliable.

Table 4: **Observation-anchor ablation in ManiSkill.** Mean ± SEM over five seeds.

	Preference-conditioned behavior				Reference stability
Reference construction	**Pref. acc. ↑**	**Target ↑**	**Balanced ↑**	**Wrong target ↓**	**Ref. drift ↓**
Frozen latent	0.793± 0.016	0.592± 0.078	0.408± 0.152	0.029± 0.029	**0.000± 0.000**
Moving latent	0.828± 0.017	0.000± 0.000	0.000± 0.000	0.004± 0.004	1.101± 0.071
Observation history (ours)	**0.865± 0.014**	**0.962± 0.034**	**0.922± 0.053**	**0.000± 0.000**	**0.000± 0.000**

reshape the state consumed by the actor; Figure 1 (a) shows that this advantage persists throughout optimization. Changing only the earlier request therefore changes the later decision in the same physical layout, directly realizing the behavioral definition of preference sufficiency.

Preference-Grounded State Construction Scales to Continuous Control. Table 3 establishes the same effect in ManiSkill. OA-LSA-DPO substantially outperforms every direct and reward-model baseline in target success and worst-profile reliability, while suppressing wrong-object behavior and consistently reversing its decision with the request. Its latent state also exhibits the strongest preference separation. The joint result is important: a reward model may predict which trajectory is preferred without ensuring that the actor-facing state retains the preference at the moment of contact. OA-LSA-DPO instead accumulates reference-relative latent advantages along the trajectory and propagates this supervision directly into the preference-facing state. Preference information is thus organized into a decision variable that determines which object is approached and displaced. Figure 1 (b)-(d) further shows that this advantage emerges during optimization, persists to convergence, and holds under both preference profiles rather than reflecting a favorable checkpoint or fixed target bias.

Why Observation Anchoring Matters. Table 4 isolates the reference construction. A moving latent reference predicts preferences but drifts with the representation and fails at control because both sides of the comparison change. A frozen latent reference avoids drift but remains tied to a state that has already discarded preference-relevant distinctions, yielding only partial control. The observation-history anchor combines a fixed comparison point with a trainable current representation, achieving the strongest preference prediction, balanced control, and zero drift. Together with Table 1, this connects non-cancelling gradients to preference-sufficient state construction.

Across the gradient test, MiniGrid, and ManiSkill, OA-LSA-DPO consistently converts preference information into decision-relevant latent structure and preference-conditioned behavior. Frozen representations cannot recover preference distinctions that have already been discarded, while learning a strong preference predictor alone does not guarantee reliable control. Additional representation visualizations, outcome-confusion analyses, counterfactual trajectories, 2D/3D manipulation geometry, score and architecture ablations, and hyperparameter sensitivity results are provided in Appendix G.

6 Conclusion

We introduced OA-LSA-DPO, a closed-form direct preference framework that jointly learns a history-dependent latent decision state and policy from trajectory comparisons. Observation-space anchoring preserves representation-level preference gradients without a reward or value model or latent-dynamics likelihood ratios. Our theory decomposes policy-matching error into actor and representation components, establishes the irreducible cost of preference-relevant latent aliasing, and explains gradient cancellation under co-encoded references. Controlled tests in MiniGrid and ManiSkill show that OA-LSA-DPO converts trajectory preferences into preference-sufficient latent structure and selective behavior. Thus, alignment in latent sequential control requires learning not only which actions to take, but also which distinctions the state must preserve.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking and language polishing. The authors reviewed all AI-assisted edits and take full responsibility for the final content of the paper.

REFERENCES

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo Perez, and Diederik M Roijers. The wasserstein believer: Learning belief updates for partially observable environments through reliable latent space models. In *International Conference on Learning Representations*, volume 2024, pp. 47537–47576, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International conference on artificial intelligence and statistics*, pp. 4447–4455. PMLR, 2024.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, et al. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. *Advances in Neural Information Processing Systems*, 37:27463–27489, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Roger R Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

Juncheng Dong, Bowen He, Moyang Guo, Ethan X Fang, Zhuoran Yang, and Vahid Tarokh. Learning in context, guided by choice: A reward-free paradigm for reinforcement learning with transformers. *arXiv preprint arXiv:2602.08244*, 2026.

Mintu Dutta, Ritesh Vyas, and Mohendra Roy. From alignment to prediction: A study of self-supervised learning and predictive representation learning. *arXiv preprint arXiv:2604.13518*, 2026.

Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. *arXiv preprint arXiv:1207.4114*, 2012.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International conference on machine learning*, pp. 2170–2179. PMLR, 2019.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018b.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

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

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, volume 45, pp. 141, 2015.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 18770–18798, 2024.

Minyoung Hwang, Luca Weihs, Chanwoo Park, Kimin Lee, Aniruddha Kembhavi, and Kiana Ehsani. Promptable behaviors: Personalizing multi-objective rewards from human preferences. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16216–16226. IEEE, 2024.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International conference on machine learning*, pp. 2117–2126. PMLR, 2018.

Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, and Donglin Wang. Beyond reward: Offline preference-guided policy optimization. *arXiv preprint arXiv:2305.16217*, 2023.

Changyeon Kim, Jongjin Park, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. Preference transformer: Modeling human preferences using transformers for rl. *arXiv preprint arXiv:2303.00957*, 2023.

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

Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in neural information processing systems*, 33:741–752, 2020.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Ian Lenz, Ross A Knepper, and Ashutosh Saxena. Deepmpc: Learning deep latent features for model predictive control. In *Robotics: Science and Systems*, volume 10, pp. 25. Rome, Italy, 2015.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*, 2022.

Zhengyu Liu, Jingliang Duan, Wenxuan Wang, Shengbo Eben Li, Yuming Yin, Ziyu Lin, and Bo Cheng. Recurrent model predictive control: Learning an explicit recurrent controller for nonlinear systems. *IEEE Transactions on Industrial Electronics*, 69(10):10437–10446, 2022.

Dylan P Losey, Hong Jun Jeon, Mengxi Li, Krishnan Srinivasan, Ajay Mandlekar, Animesh Garg, Jeannette Bohg, and Dorsa Sadigh. Learning latent actions to control assistive robots. *Autonomous robots*, 46(1):115–147, 2022.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural networks*, 152:267–275, 2022.

Jiahao Mei, Xuenan Xu, Zeyu Xie, Zihao Zheng, Ye Tao, Yue Ding, and Mengyue Wu. Lara-gen: Enabling continuous emotion control for music generation models via latent affective representation alignment. *arXiv preprint arXiv:2510.05875*, 2025.

Katherine Metcalf, Miguel Sarabia, Natalie Mackraz, and Barry-John Theobald. Sample-efficient preference-based reinforcement learning with dynamics aware rewards. *arXiv preprint arXiv:2402.17975*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. *arXiv preprint arXiv:2203.10050*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature communications*, 12(1):1417, 2021.

Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. Despot: Online pomdp planning with regularization. *Advances in neural information processing systems*, 26, 2013.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Jayden Teoh, Manan Tomar, Kwangjun Ahn, Edward S Hu, Tim Pearce, Pratyusha Sharma, Akshay Krishnamurthy, Riashat Islam, Alex Lamb, and John Langford. Next-latent prediction transformers learn compact world models. *arXiv preprint arXiv:2511.05963*, 2025.

Richa Verma, Bavish Kulur, Sanjay Chawla, and Balaraman Ravindran. Prefine: Preference-based implicit reward and cost fine-tuning for safety alignment. *arXiv preprint arXiv:2605.21225*, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

Sen Wang, Sanping Zhou, Huaiyi Dong, Kun Xia, Gang Hua, and Le Wang. Action-controlled scale-wise flow matching for embodied world models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jun-Jie Yang, Chia-Heng Hsu, Kui-Yuan Chen, and Ping-Chun Hsieh. From reward-free representations to preferences: Rethinking offline preference-based reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026.

Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. Beyond bradley-terry models: A general preference model for language model alignment. *arXiv preprint arXiv:2410.02197*, 2024.

Guangyu Zhao, Kewei Lian, Haowei Lin, Haobo Fu, Qiang Fu, Shaofei Cai, Zihao Wang, and Yitao Liang. Optimizing latent goal by learning from trajectory preference. 2024.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40277–40290, 2026.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Related Work

A.1 Latent Representations for Sequential Control

Learning a compact decision state is central to sequential control under partial observability. Recurrent policies and belief-state methods summarize observation–action histories to support action selection (Hausknecht & Stone, 2015; Igl et al., 2018). Latent-dynamics and world-model approaches further learn internal states from reconstruction, future prediction, or imagined control objectives (Ha & Schmidhuber, 2018b; Hafner et al., 2019a;b; Lee et al., 2020). Related work on state abstraction, bisimulation, and control-oriented representation learning seeks latent states that preserve transition, reward, or value information relevant to a specified task (Li et al., 2006; Ferns et al., 2012; Zhang et al., 2020; Gelada et al., 2019). These approaches have established that a learned state need not reconstruct the complete observation history, provided that it retains the information required by the associated prediction or control objective. However, their representations are primarily supervised by environment dynamics, reconstruction targets, or externally specified rewards. They do not require histories mapped to the same latent state to induce the same preference-optimal action distribution, nor do they use trajectory comparisons to directly shape the state consumed by the actor. Consequently, a representation that is sufficient for prediction or reward-based control may remain insufficient for preference-aligned decision making, and this representational deficiency cannot generally be corrected by optimizing the actor alone.

A.2 Preference Optimization

Preference optimization provides a general mechanism for aligning learned models when desired behavior is difficult to express through an explicit scalar objective. Early reinforcement learning from human feedback (RLHF) methods fit a reward model from pairwise comparisons and then optimize the learned reward using reinforcement learning (Christiano et al., 2017; Ouyang et al., 2022; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). Direct preference methods instead optimize the model from preferred and rejected examples without an explicit reward-learning stage (Rafailov et al., 2023; Azar et al., 2024; Zhao et al., 2023). This paradigm has been applied broadly to language-model alignment, text-to-image and diffusion models, multimodal generation, recommendation, and embodied agents (Black et al., 2024; Wallace et al., 2024; Wang et al., 2024; Chen et al., 2024; Hwang et al., 2024). Despite their different applications, these methods generally compare complete outputs under a fixed and directly observed conditioning variable, such as a prompt, image, instruction, or user context. Their reference-relative objectives are therefore defined in a common output space whose semantics remain unchanged during optimization. Although the internal neural features of these models may be updated, the conditioning variable itself is not treated as a learned sequential decision state. General preference optimization consequently does not address how to maintain a well-defined reference comparison when the history-dependent state governing future actions is itself modified by preference gradients.

A.3 Preference Learning for Sequential Control

Preference-based reinforcement learning extends pairwise supervision to temporally extended behavior, typically by collecting comparisons between trajectory segments. Most methods first infer a scalar reward or preference model and subsequently optimize a policy against that model (Christiano et al., 2017; Lee et al., 2021; Park et al., 2022). Later approaches improve feedback efficiency through uncertainty, data augmentation, non-Markovian preference models, or dynamics-aware representations (Liang et al., 2022; Kim et al., 2023; Metcalf et al., 2024). These methods may learn useful representations, but those representations primarily support reward estimation, leaving policy optimization as a separate stage. Direct sequential preference methods remove the explicit reward model and optimize policies from trajectory comparisons, but they operate on supplied environment states or fixed policy representations (Hejna et al., 2024; Verma et al., 2026; Dong et al., 2026). Other approaches infer global task contexts, tune latent goals, or adapt reward-free representations from preferences (Kang et al., 2023; Zhao et al., 2024; Yang et al., 2026). Such latent variables are typically appended to an observed state, optimized while the policy backbone is fixed, or learned through auxiliary context and representation objectives. They are therefore different from the recurrent policy-facing state through which every action is generated. None of these approaches provides a closed-form, reference-relative preference objective that jointly learns an explicit history-dependent

latent decision state and the actor operating exclusively upon it. OA-LSA-DPO closes this gap by anchoring the reference policy on observable histories while allowing trajectory preferences to directly construct the latent state used for sequential action selection.

B Trajectory-Ratio Cancellation and the Latent-Marginal Obstruction

This appendix distinguishes the trajectory distribution induced by a policy in a fixed environment from the likelihood assigned by a learned latent generative model. The former admits cancellation of shared environment factors, whereas the latter generally does not reduce to an action-only likelihood ratio when its representation is learnable.

Cancellation on the observable-history process. For a partially observable environment, define the environment-induced causal likelihood

$$\kappa_P\left(o_{0:H} \,\|\, a_{0:H-1}\right) := \int p_0(s_0) O(o_0 \mid s_0) \prod_{t=0}^{H-1} P(s_{t+1} \mid s_t, a_t) O(o_{t+1} \mid s_{t+1}) \, \mathrm{d}s_{0:H}. \tag{19}$$

A history-dependent policy $\pi(a_t \mid h_t^o)$ therefore induces the observable trajectory density

$$p_\pi(\tau^o) = \kappa_P\left(o_{0:H} \,\|\, a_{0:H-1}\right) \prod_{t=0}^{H-1} \pi(a_t \mid h_t^o). \tag{20}$$

If π and a reference policy μ_{ref} interact with the same initial-state distribution, transition kernel, and emission kernel, then, under the usual support condition,

$$\begin{aligned} \log \frac{p_\pi(\tau^o)}{p_{\mu_{\mathrm{ref}}}(\tau^o)} &= \log \frac{\kappa_P\left(o_{0:H} \,\|\, a_{0:H-1}\right)}{\kappa_P\left(o_{0:H} \,\|\, a_{0:H-1}\right)} + \sum_{t=0}^{H-1} \log \frac{\pi(a_t \mid h_t^o)}{\mu_{\mathrm{ref}}(a_t \mid h_t^o)} \\ &= \sum_{t=0}^{H-1} \log \frac{\pi(a_t \mid h_t^o)}{\mu_{\mathrm{ref}}(a_t \mid h_t^o)}. \end{aligned} \tag{21}$$

Thus, partial observability does not prevent cancellation when both policies are defined on the same observable-history process. The environment likelihood is shared and independent of the policy and representation parameters.

Likelihood ratio of a learned latent generative process. Consider instead a generic latent-model likelihood. Conditioning on the initial observation o_0, its joint density may be written as

$$p^{\mathrm{lat}}_{\eta,\rho,\theta}\left(\tau^o, z_{0:H}\right) = q_\eta(z_0 \mid o_0) \prod_{t=0}^{H-1} \Big[\pi_\theta(a_t \mid z_t) \, p^z_\rho\left(z_{t+1} \mid z_t, a_t\right) p^o_\rho\left(o_{t+1} \mid z_{t+1}\right) \Big], \tag{22}$$

where q_η initializes or infers the latent state, p^z_ρ is the learned latent transition model, and p^o_ρ is the observation model. This construction is introduced only to analyze a naive latent-trajectory extension of DPO; it is not the likelihood optimized by our method.

Let $(\bar{\eta}, \bar{\rho}, \bar{\theta})$ denote the corresponding reference parameters. Suppose temporarily that the two models use a common latent coordinate system and satisfy the required absolute-continuity conditions. Define the joint log-density ratio

$$R(\tau^o, z_{0:H}) := \log \frac{p^{\mathrm{lat}}_{\eta,\rho,\theta}\left(\tau^o, z_{0:H}\right)}{p^{\mathrm{lat}}_{\bar{\eta},\bar{\rho},\bar{\theta}}\left(\tau^o, z_{0:H}\right)}. \tag{23}$$

Substituting Eq. (22) into Eq. (23) gives

$$\begin{aligned} R(\tau^o, z_{0:H}) = \log \frac{q_\eta(z_0 \mid o_0)}{q_{\bar{\eta}}(z_0 \mid o_0)} &+ \sum_{t=0}^{H-1} \log \frac{\pi_\theta(a_t \mid z_t)}{\pi_{\bar{\theta}}(a_t \mid z_t)} + \sum_{t=0}^{H-1} \log \frac{p^z_\rho\left(z_{t+1} \mid z_t, a_t\right)}{p^z_{\bar{\rho}}\left(z_{t+1} \mid z_t, a_t\right)} \\ &+ \sum_{t=0}^{H-1} \log \frac{p^o_\rho\left(o_{t+1} \mid z_{t+1}\right)}{p^o_{\bar{\rho}}\left(o_{t+1} \mid z_{t+1}\right)}. \end{aligned} \tag{24}$$

Algorithm 1 Observation-Anchored Latent Soft-Advantage DPO

Require: Trajectory data $\mathcal{D}_{\mathrm{traj}}$, preference pairs $\mathcal{D}_{\mathrm{pref}}$, coefficient β, and weight λ_{pref}.

1: Train or initialize the latent sequential model (f_η, ρ) using $\mathcal{L}_{\mathrm{latent}}$ on $\mathcal{D}_{\mathrm{traj}}$.
2: Train and freeze the observation-history reference policy $\mu_{\mathrm{ref}}(a \mid h^o)$, or cache its action log-likelihoods.
3: Initialize the latent actor $\pi_\theta(a \mid z)$.
4: **for** each joint update **do**
5: Sample a trajectory minibatch $\mathcal{B}_{\mathrm{traj}}$ from $\mathcal{D}_{\mathrm{traj}}$.
6: Sample a preference minibatch $\mathcal{B}_{\mathrm{pref}}$ from $\mathcal{D}_{\mathrm{pref}}$.
7: Evaluate $\mathcal{L}_{\mathrm{latent}}(\eta, \rho)$ on $\mathcal{B}_{\mathrm{traj}}$.
8: For each rollout in $\mathcal{B}_{\mathrm{pref}}$, reconstruct h_t^o and compute $z_t = f_\eta(h_t^o)$.
9: Compute the observation-anchored soft advantages using Eq. (9) and accumulate the trajectory scores using Eq. (11).
10: Evaluate $\mathcal{L}_{\text{OA-LSA-DPO}}(\eta, \theta)$ using Eq. (12).
11: Update (η, ρ, θ) using the joint objective in Eq. (13).
12: **end for**π_θ.

The ratio therefore contains latent-initialization, policy, transition, and observation terms. The non-policy terms cancel pathwise only under restrictive equality conditions such as $q_\eta = q_{\bar\eta}$, $p_\rho^z = p_{\bar\rho}^z$, and $p_\rho^o = p_{\bar\rho}^o$. These conditions are generally not preserved when preference learning updates the representation or predictive model.

More fundamentally, the supervisor observes τ^o, not $z_{0:H}$. The relevant likelihood is therefore the observable marginal

$$p^{\mathrm{lat}}_{\eta,\rho,\theta}(\tau^o) = \int p^{\mathrm{lat}}_{\eta,\rho,\theta}\left(\tau^o, z_{0:H}\right) \mathrm{d}z_{0:H}. \tag{25}$$

Let $p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}\left(z_{0:H} \mid \tau^o\right)$ denote the reference posterior over latent paths. Using Eq. (23), we obtain

$$\begin{aligned}\frac{p^{\mathrm{lat}}_{\eta,\rho,\theta}(\tau^o)}{p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}(\tau^o)} &= \int p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}\left(z_{0:H} \mid \tau^o\right) \frac{p^{\mathrm{lat}}_{\eta,\rho,\theta}\left(\tau^o, z_{0:H}\right)}{p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}\left(\tau^o, z_{0:H}\right)} \mathrm{d}z_{0:H} \\ &= \mathbb{E}_{z_{0:H}\sim p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}(\cdot\mid\tau^o)}\left[\exp\left(R(\tau^o, z_{0:H})\right)\right].\end{aligned} \tag{26}$$

Taking logarithms yields the exact identity

$$\log \frac{p^{\mathrm{lat}}_{\eta,\rho,\theta}(\tau^o)}{p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}(\tau^o)} = \log \mathbb{E}_{z_{0:H}\sim p^{\mathrm{lat}}_{\bar\eta,\bar\rho,\bar\theta}(\cdot\mid\tau^o)}\left[\exp\left(R(\tau^o, z_{0:H})\right)\right]. \tag{27}$$

The right-hand side is a log-moment-generating expression over complete latent paths. It is not, in general, equal to a pointwise sum of action log-ratios. Such a reduction requires both that the non-policy terms in Eq. (24) cancel and that the remaining action-ratio term be evaluated on a shared deterministic latent path or be almost surely constant under the reference posterior.

Finally, the pointwise ratio in Eq. (24) presupposes an identification between the current and reference latent coordinates. When $z_t = f_\eta(h_t^o)$ and $\bar z_t = f_{\bar\eta}(h_t^o)$, the two encoders may assign different coordinates to the same observable history. A comparison formed directly in the learned latent space is therefore representation-dependent and is not invariant to general latent reparameterizations.

Consequently, a naive likelihood ratio over a learned latent generative process does not inherit the action-only factorization of standard DPO. Our method instead defines the learned and reference policies on the common observable-history process in Eq. (20). The shared physical environment then cancels according to Eq. (21), while preference gradients remain free to update the latent state used by the current actor.

C Derivation of the Observation-Anchored Soft-Advantage Identity

This section provides the complete derivation of Equations Eq. (6)–Eq. (8). We consider the history MDP introduced in Section 3.1, with history state $h^o \in \mathcal{H}$, action $a \in \mathcal{A}$, environment transition kernel $P_{\text{env}}(h^{o\prime} \mid h^o, a)$, and fixed reference policy $\mu_{\text{ref}}(a \mid h^o)$.

We assume that $\mu_{\text{ref}}(a \mid h^o) > 0$ on the support of every policy under consideration and that the relevant log-partition functions are finite. These conditions ensure that the KL divergence and the policy ratios below are well defined. We present the derivation for continuous action spaces; integrals are replaced by sums in the discrete case.

Proposition C.1 (Observation-anchored soft-control identity)**.** *For a fixed history h^o, consider the statewise KL-regularized Bellman optimization*

$$V_\beta^\star(h^o) = \max_{\pi(\cdot|h^o)} \int_{\mathcal{A}} \pi(a \mid h^o) \left[Q_\beta^\star(h^o, a) - \beta \log \frac{\pi(a \mid h^o)}{\mu_{\text{ref}}(a \mid h^o)} \right] \mathrm{d}a, \tag{28}$$

subject to

$$\pi(a \mid h^o) \geq 0, \qquad \int_{\mathcal{A}} \pi(a \mid h^o)\, \mathrm{d}a = 1. \tag{29}$$

Then the unique optimal policy is

$$\pi_\beta^\star(a \mid h^o) = \mu_{\text{ref}}(a \mid h^o) \exp\left(\frac{Q_\beta^\star(h^o, a) - V_\beta^\star(h^o)}{\beta} \right), \tag{30}$$

where $V_\beta^\star(h^o) = \beta \log \int_{\mathcal{A}} \mu_{\text{ref}}(a \mid h^o) \exp\left(\frac{Q_\beta^\star(h^o,a)}{\beta}\right) \mathrm{d}a$. Consequently, the optimal soft advantage $A_\beta^\star(h^o, a) := Q_\beta^\star(h^o, a) - V_\beta^\star(h^o)$ satisfies

$$A_\beta^\star(h^o, a) = \beta \log \frac{\pi_\beta^\star(a \mid h^o)}{\mu_{\text{ref}}(a \mid h^o)}. \tag{31}$$

Proof. Fix an arbitrary history h^o. To simplify notation within the proof, let $p(a) := \pi(a \mid h^o)$, $m(a) := \mu_{\text{ref}}(a \mid h^o)$, and $q(a) := Q_\beta^\star(h^o, a)$. The statewise optimization in Eq. (28) becomes

$$\max_{p \in \Delta(\mathcal{A})} \left\{ \int_{\mathcal{A}} p(a) q(a)\, \mathrm{d}a - \beta \int_{\mathcal{A}} p(a) \log \frac{p(a)}{m(a)}\, \mathrm{d}a \right\}, \tag{32}$$

where $\Delta(\mathcal{A})$ denotes the set of normalized action densities. The first term is linear in p, while the negative KL term is strictly concave in p for $\beta > 0$. Therefore, any feasible stationary point is the unique global maximizer.

Introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ for the normalization constraint. The Lagrangian is

$$\mathcal{J}(p, \lambda) = \int_{\mathcal{A}} p(a) q(a)\, \mathrm{d}a - \beta \int_{\mathcal{A}} p(a) \log \frac{p(a)}{m(a)}\, \mathrm{d}a + \lambda \left(1 - \int_{\mathcal{A}} p(a)\, \mathrm{d}a \right). \tag{33}$$

For any action a, the functional derivative with respect to $p(a)$ is

$$\frac{\delta \mathcal{J}}{\delta p(a)} = q(a) - \beta \left(\log \frac{p(a)}{m(a)} + 1 \right) - \lambda. \tag{34}$$

At the optimum, the stationarity condition gives

$$q(a) - \beta \left(\log \frac{p^\star(a)}{m(a)} + 1 \right) - \lambda = 0. \tag{35}$$

Rearranging,

$$\log \frac{p^\star(a)}{m(a)} = \frac{q(a)}{\beta} - 1 - \frac{\lambda}{\beta}. \tag{36}$$

Exponentiating both sides yields

$$p^\star(a) = m(a) \exp\left(\frac{q(a)}{\beta} \right) \exp\left(-1 - \frac{\lambda}{\beta} \right). \tag{37}$$

The last exponential factor does not depend on a and is therefore determined by normalization. Define

$$Z(h^o) := \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_\beta(h^o, a)}{\beta}\right) \mathrm{d}a. \tag{38}$$

Imposing $\int_{\mathcal{A}} p^\star(a)\,\mathrm{d}a = 1$ in Eq. (37) gives $\exp\left(-1 - \frac{\lambda}{\beta}\right) = \frac{1}{Z(h^o)}$. Consequently,

$$\pi^\star_\beta(a \mid h^o) = \frac{\mu_{\mathrm{ref}}(a \mid h^o) \exp\left(Q^\star_\beta(h^o, a)/\beta\right)}{Z(h^o)}. \tag{39}$$

We next evaluate the optimal Bellman objective. From Eq. (39),

$$\log \frac{\pi^\star_\beta(a \mid h^o)}{\mu_{\mathrm{ref}}(a \mid h^o)} = \frac{Q^\star_\beta(h^o, a)}{\beta} - \log Z(h^o). \tag{40}$$

Substituting this identity into Eq. (28) gives

$$\begin{aligned} V^\star_\beta(h^o) &= \int_{\mathcal{A}} \pi^\star_\beta(a \mid h^o) \left[Q^\star_\beta(h^o, a) - \beta\left(\frac{Q^\star_\beta(h^o, a)}{\beta} - \log Z(h^o)\right)\right] \mathrm{d}a \\ &= \int_{\mathcal{A}} \pi^\star_\beta(a \mid h^o) \beta \log Z(h^o)\,\mathrm{d}a \\ &= \beta \log Z(h^o). \end{aligned} \tag{41}$$

Combining Eq. (41) with Eq. (38) yields

$$V^\star_\beta(h^o) = \beta \log \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_\beta(h^o, a)}{\beta}\right) \mathrm{d}a. \tag{42}$$

Using $Z(h^o) = \exp(V^\star_\beta(h^o)/\beta)$ in Eq. (39) gives

$$\begin{aligned} \pi^\star_\beta(a \mid h^o) &= \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_\beta(h^o, a)}{\beta} - \frac{V^\star_\beta(h^o)}{\beta}\right) \\ &= \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_\beta(h^o, a) - V^\star_\beta(h^o)}{\beta}\right), \end{aligned} \tag{43}$$

which proves Eq. (30).

For completeness, global optimality can also be verified directly through a KL decomposition. For any feasible policy $\pi(\cdot \mid h^o)$, Equation Eq. (39) implies

$$\int_{\mathcal{A}} \pi(a \mid h^o) \left[Q^\star_\beta(h^o, a) - \beta \log \frac{\pi(a \mid h^o)}{\mu_{\mathrm{ref}}(a \mid h^o)}\right] \mathrm{d}a = \beta \log Z(h^o) - \beta D_{\mathrm{KL}}\left(\pi(\cdot \mid h^o) \,\|\, \pi^\star_\beta(\cdot \mid h^o)\right). \tag{44}$$

Since the KL divergence is nonnegative, the objective is maximized uniquely at $\pi = \pi^\star_\beta$, and its maximum is $\beta \log Z(h^o)$.

Finally, taking logarithms of Eq. (43) gives $\beta \log \frac{\pi^\star_\beta(a|h^o)}{\mu_{\mathrm{ref}}(a|h^o)} = Q^\star_\beta(h^o, a) - V^\star_\beta(h^o)$. By the definition of $A^\star_\beta$, this is exactly $A^\star_\beta(h^o, a) = \beta \log \frac{\pi^\star_\beta(a|h^o)}{\mu_{\mathrm{ref}}(a|h^o)}$, which proves Eq. (31). □

D Proofs for Sequential Reward Equivalence and Canonical Latent Advantage

We prove Proposition 4.1 and Corollary 4.2. We assume either an infinite-horizon discounted process with $\gamma \in [0, 1)$ and bounded rewards, or a finite-horizon process with an absorbing terminal history and zero terminal potential. The reference policy has full support over the admissible action set.

For a reward r, the optimal regularized Bellman equations are

$$Q^\star_{\beta,r}(h^o, a) = r(h^o, a) + \gamma \mathbb{E}_{h^{o\prime} \sim P(\cdot|h^o,a)}\left[V^\star_{\beta,r}(h^{o\prime})\right], \tag{45}$$

$$V^\star_{\beta,r}(h^o) = \beta \log \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_{\beta,r}(h^o,a)}{\beta}\right) \mathrm{d}a, \tag{46}$$

$$\pi^\star_{\beta,r}(a \mid h^o) = \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_{\beta,r}(h^o,a) - V^\star_{\beta,r}(h^o)}{\beta}\right). \tag{47}$$

Consequently,

$$A^\star_{\beta,r}(h^o,a) = \beta \log \frac{\pi^\star_{\beta,r}(a \mid h^o)}{\mu_{\mathrm{ref}}(a \mid h^o)}. \tag{48}$$

Proof of Proposition 4.1. We prove the equivalence in three steps.

Policy equivalence and advantage equivalence. If $r \simeq_\beta \widetilde{r}$, then $\pi^\star_{\beta,r}(a \mid h^o) = \pi^\star_{\beta,\widetilde{r}}(a \mid h^o)$ for every (h^o, a). Applying Eq. (48) to both rewards gives $A^\star_{\beta,r}(h^o,a) = A^\star_{\beta,\widetilde{r}}(h^o,a)$. Conversely, equality of the advantages implies equality of the policies through Eq. (47). Hence, statements 1 and 2 are equivalent.

Potential shaping implies advantage equivalence. Suppose $\widetilde{r}(h^o,a) = r(h^o,a) + \gamma\mathbb{E}\left[\Psi(h^{o\prime}) \mid h^o, a\right] - \Psi(h^o)$. Define $\widetilde{Q}(h^o,a) := Q^\star_{\beta,r}(h^o,a) - \Psi(h^o)$ and $\widetilde{V}(h^o) := V^\star_{\beta,r}(h^o) - \Psi(h^o)$. Then

$$\begin{aligned}
&\widetilde{r}(h^o,a) + \gamma\mathbb{E}\left[\widetilde{V}(h^{o\prime}) \mid h^o, a\right] \\
&= r(h^o,a) + \gamma\mathbb{E}\left[\Psi(h^{o\prime}) \mid h^o, a\right] - \Psi(h^o) + \gamma\mathbb{E}\left[V^\star_{\beta,r}(h^{o\prime}) - \Psi(h^{o\prime}) \mid h^o, a\right] \\
&= Q^\star_{\beta,r}(h^o,a) - \Psi(h^o) = \widetilde{Q}(h^o,a).
\end{aligned} \tag{49}$$

Moreover,

$$\begin{aligned}
&\beta \log \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{\widetilde{Q}(h^o,a)}{\beta}\right) \mathrm{d}a \\
&= \beta \log \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\star_{\beta,r}(h^o,a) - \Psi(h^o)}{\beta}\right) \mathrm{d}a \\
&= V^\star_{\beta,r}(h^o) - \Psi(h^o) = \widetilde{V}(h^o).
\end{aligned} \tag{50}$$

Thus, $(\widetilde{Q}, \widetilde{V})$ satisfies the optimal soft Bellman equations for $\widetilde{r}$. By uniqueness of the Bellman solution, we have $Q^\star_{\beta,\widetilde{r}} = Q^\star_{\beta,r} - \Psi$ and $V^\star_{\beta,\widetilde{r}} = V^\star_{\beta,r} - \Psi$. Therefore,

$$\begin{aligned}
A^\star_{\beta,\widetilde{r}}(h^o,a) &= Q^\star_{\beta,\widetilde{r}}(h^o,a) - V^\star_{\beta,\widetilde{r}}(h^o) \\
&= Q^\star_{\beta,r}(h^o,a) - V^\star_{\beta,r}(h^o) \\
&= A^\star_{\beta,r}(h^o,a).
\end{aligned} \tag{51}$$

Advantage equivalence implies potential shaping. Suppose $A^\star_{\beta,\widetilde{r}}(h^o,a) = A^\star_{\beta,r}(h^o,a)$ for every (h^o,a), and define $\Psi(h^o) := V^\star_{\beta,r}(h^o) - V^\star_{\beta,\widetilde{r}}(h^o)$. Advantage equality gives

$$Q^\star_{\beta,\widetilde{r}}(h^o,a) - Q^\star_{\beta,r}(h^o,a) = -\Psi(h^o). \tag{52}$$

Using the Bellman equations,

$$\begin{aligned}
\widetilde{r}(h^o,a) - r(h^o,a) &= Q^\star_{\beta,\widetilde{r}}(h^o,a) - Q^\star_{\beta,r}(h^o,a) \\
&\quad - \gamma\mathbb{E}\left[V^\star_{\beta,\widetilde{r}}(h^{o\prime}) - V^\star_{\beta,r}(h^{o\prime}) \mid h^o, a\right] \\
&= -\Psi(h^o) + \gamma\mathbb{E}\left[\Psi(h^{o\prime}) \mid h^o, a\right],
\end{aligned} \tag{53}$$

which proves Eq. (14).

Canonical representative. Choose $\Psi(h^o) = V^\star_{\beta,r}(h^o)$. The corresponding shaped reward is

$$\begin{aligned} r^\circ(h^o, a) &= r(h^o, a) + \gamma \mathbb{E}\left[V^\star_{\beta,r}(h^{o\prime}) \mid h^o, a\right] - V^\star_{\beta,r}(h^o) \\ &= Q^\star_{\beta,r}(h^o, a) - V^\star_{\beta,r}(h^o) \\ &= A^\star_{\beta,r}(h^o, a). \end{aligned} \tag{54}$$

Then we have $V^\star_{\beta,r^\circ}(h^o) = 0$ and $Q^\star_{\beta,r^\circ}(h^o, a) = A^\star_{\beta,r}(h^o, a)$.

To prove uniqueness, let $\widehat{r}$ be control-equivalent to r and satisfy $V^\star_{\beta,\widehat{r}}(h^o) = 0$ for every history. Control equivalence implies $A^\star_{\beta,\widehat{r}} = A^\star_{\beta,r}$. It follows that $Q^\star_{\beta,\widehat{r}} = A^\star_{\beta,\widehat{r}} + V^\star_{\beta,\widehat{r}} = A^\star_{\beta,r}$. Using the Bellman equation once more,

$$\begin{aligned} \widehat{r}(h^o, a) &= Q^\star_{\beta,\widehat{r}}(h^o, a) - \gamma \mathbb{E}\left[V^\star_{\beta,\widehat{r}}(h^{o\prime}) \mid h^o, a\right] \\ &= A^\star_{\beta,r}(h^o, a) = r^\circ(h^o, a). \end{aligned} \tag{55}$$

Hence, the zero-value representative is unique. □

Proof of Corollary 4.2. Let $r^\circ_{\eta,\theta}(h^o, a) = A_{\eta,\theta}(h^o, a) = \beta \log \frac{\pi_{\eta,\theta}(a|h^o)}{\mu_{\mathrm{ref}}(a|h^o)}$. Consider $Q^\circ(h^o, a) := A_{\eta,\theta}(h^o, a)$ and $V^\circ(h^o) := 0$. The action-value Bellman equation holds because

$$r^\circ_{\eta,\theta}(h^o, a) + \gamma \mathbb{E}\left[V^\circ(h^{o\prime}) \mid h^o, a\right] = A_{\eta,\theta}(h^o, a) = Q^\circ(h^o, a). \tag{56}$$

The soft value equation also holds:

$$\begin{aligned} &\beta \log \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\circ(h^o, a)}{\beta}\right) \mathrm{d}a \\ &= \beta \log \int_{\mathcal{A}} \mu_{\mathrm{ref}}(a \mid h^o) \frac{\pi_{\eta,\theta}(a \mid h^o)}{\mu_{\mathrm{ref}}(a \mid h^o)} \mathrm{d}a \\ &= \beta \log \int_{\mathcal{A}} \pi_{\eta,\theta}(a \mid h^o) \mathrm{d}a \\ &= 0 = V^\circ(h^o). \end{aligned} \tag{57}$$

Therefore, (Q°, V°) satisfies the optimal regularized Bellman equations for $r^\circ_{\eta,\theta}$. The corresponding optimal policy is

$$\begin{aligned} \pi^\star_{\beta, r^\circ_{\eta,\theta}}(a \mid h^o) &= \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\frac{Q^\circ(h^o, a) - V^\circ(h^o)}{\beta}\right) \\ &= \mu_{\mathrm{ref}}(a \mid h^o) \exp\left(\log \frac{\pi_{\eta,\theta}(a \mid h^o)}{\mu_{\mathrm{ref}}(a \mid h^o)}\right) \\ &= \pi_{\eta,\theta}(a \mid h^o). \end{aligned} \tag{58}$$

This proves all three identities in Corollary 4.2. □

E Proofs for the Exact Decomposition of Latent Preference Error

We prove Theorem 4.3 and Corollary 4.4. We assume that the history and action spaces are standard Borel spaces, so that the conditional distribution $\nu(\mathrm{d}h^o \mid z)$ exists. All policies admit strictly positive densities with respect to a common base measure on $\mathcal{A}$.

Proof of Theorem 4.3. Fix a reachable latent state z and abbreviate $m_z(a) := \mathbb{E}_\nu\left[\log p_{h^o}(a) \mid f_\eta(h^o) = z\right]$. For any latent action distribution $q(\cdot \mid z)$, its conditional preference error is

$$\begin{aligned} &\mathbb{E}_\nu\left[D_{\mathrm{KL}}\left(q(\cdot \mid z) \,\|\, p_{h^o}\right) \mid f_\eta(h^o) = z\right] \\ &= \mathbb{E}_\nu\left[\int_{\mathcal{A}} q(a \mid z) \log \frac{q(a \mid z)}{p_{h^o}(a)} \mathrm{d}a \,\middle|\, f_\eta(h^o) = z\right] \end{aligned}$$

$$= \int_{\mathcal{A}} q(a \mid z) \left[\log q(a \mid z) - m_z(a)\right] \mathrm{d}a. \tag{59}$$

By Eq. (15), $\log q^{\dagger}_{\eta,\nu}(a \mid z) = m_z(a) - \log Z^{\nu}_{\eta}(z)$, or equivalently, $m_z(a) = \log q^{\dagger}_{\eta,\nu}(a \mid z) + \log Z^{\nu}_{\eta}(z)$. Substituting this into Eq. (59) gives

$$\begin{aligned}
&\mathbb{E}_{\nu}\left[D_{\mathrm{KL}}\left(q(\cdot \mid z) \,\|\, p_{h^o}\right) \mid f_{\eta}(h^o) = z\right] \\
&= \int_{\mathcal{A}} q(a \mid z) \log \frac{q(a \mid z)}{q^{\dagger}_{\eta,\nu}(a \mid z)} \mathrm{d}a - \log Z^{\nu}_{\eta}(z) \int_{\mathcal{A}} q(a \mid z) \mathrm{d}a \\
&= D_{\mathrm{KL}}\left(q(\cdot \mid z) \,\|\, q^{\dagger}_{\eta,\nu}(\cdot \mid z)\right) - \log Z^{\nu}_{\eta}(z).
\end{aligned} \tag{60}$$

Multiplying by β and taking expectation over $z \sim \nu_{\eta}$ proves Eq. (16).

Because the KL divergence is nonnegative, $\mathcal{R}_{\nu}(f_{\eta}, q) \geq \beta \mathbb{E}_{z \sim \nu_{\eta}}\left[-\log Z^{\nu}_{\eta}(z)\right]$. Equality holds if and only if $q(\cdot \mid z) = q^{\dagger}_{\eta,\nu}(\cdot \mid z)$ for ν_{η}-almost every z. Therefore, $q^{\dagger}_{\eta,\nu}$ is the unique risk-minimizing latent actor, and $\inf_q \mathcal{R}_{\nu}(f_{\eta}, q) = \mathcal{E}_{\mathrm{rep}}(f_{\eta}; \nu)$.

It remains to characterize when the representation error vanishes. For every action a, Jensen's inequality gives

$$\exp\left(\mathbb{E}_{\nu}\left[\log p_{h^o}(a) \mid f_{\eta}(h^o) = z\right]\right) \leq \mathbb{E}_{\nu}\left[p_{h^o}(a) \mid f_{\eta}(h^o) = z\right]. \tag{61}$$

Integrating both sides over the action space yields

$$\begin{aligned}
Z^{\nu}_{\eta}(z) &\leq \int_{\mathcal{A}} \mathbb{E}_{\nu}\left[p_{h^o}(a) \mid f_{\eta}(h^o) = z\right] \mathrm{d}a \\
&= \mathbb{E}_{\nu}\left[\int_{\mathcal{A}} p_{h^o}(a) \mathrm{d}a \,\middle|\, f_{\eta}(h^o) = z\right] \\
&= 1.
\end{aligned} \tag{62}$$

Hence, $-\log Z^{\nu}_{\eta}(z) \geq 0$.

Suppose first that f_{η} is preference-sufficient ν-almost surely. Then, for ν_{η}-almost every z, there exists a distribution p_z such that $p_{h^o} = p_z$ for $\nu(\mathrm{d}h^o \mid z)$-almost every h^o. It follows that $m_z(a) = \log p_z(a)$ and $Z^{\nu}_{\eta}(z) = \int_{\mathcal{A}} p_z(a) \mathrm{d}a = 1$. Therefore, $\mathcal{E}_{\mathrm{rep}}(f_{\eta}; \nu) = 0$.

Conversely, suppose $\mathcal{E}_{\mathrm{rep}}(f_{\eta}; \nu) = 0$. Since $-\log Z^{\nu}_{\eta}(z) \geq 0$, its expectation can vanish only if $Z^{\nu}_{\eta}(z) = 1$ for ν_{η}-almost every z. Equality must therefore hold in Eq. (61) for almost every action. Because the logarithm is strictly concave on the positive real line, equality in Jensen's inequality implies that, for ν_{η}-almost every z and almost every a, $p_{h^o}(a)$ is constant for $\nu(\mathrm{d}h^o \mid z)$-almost every h^o. Thus, histories assigned to the same latent state have identical optimal action distributions, up to null sets. This is precisely preference sufficiency under ν, proving Eq. (15). □

Two-history latent aliasing. Suppose a latent state z contains exactly two histories h^o and $\widetilde{h}^o$ with equal conditional probability. Then

$$m^{\nu}_{\eta}(a, z) = \frac{1}{2} \log p_{h^o}(a) + \frac{1}{2} \log p_{\widetilde{h}^o}(a), \tag{63}$$

and hence

$$\exp\left(m^{\nu}_{\eta}(a, z)\right) = \sqrt{p_{h^o}(a) p_{\widetilde{h}^o}(a)}. \tag{64}$$

Therefore,

$$Z^{\nu}_{\eta}(z) = \int_{\mathcal{A}} \sqrt{p_{h^o}(a) p_{\widetilde{h}^o}(a)} \mathrm{d}a. \tag{65}$$

By the Cauchy–Schwarz inequality,

$$Z^{\nu}_{\eta}(z) \leq \sqrt{\int_{\mathcal{A}} p_{h^o}(a) \mathrm{d}a} \sqrt{\int_{\mathcal{A}} p_{\widetilde{h}^o}(a) \mathrm{d}a} = 1, \tag{66}$$

with equality if and only if $p_{h^o} = p_{\widetilde{h}^o}$ almost everywhere.

Proof of Corollary 4.4. Let the regularized value of a history policy π be

$$V_\beta^\pi(h_0^o) := \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(r^\star(h_t^o,a_t)-\beta\log\frac{\pi(a_t\mid h_t^o)}{\mu_{\mathrm{ref}}(a_t\mid h_t^o)}\right)\,\middle|\,h_0^o\right]. \tag{67}$$

The optimal soft Bellman equation gives

$$r^\star(h^o,a)+\gamma\mathbb{E}\left[V_\beta^\star(h^{o\prime})\mid h^o,a\right]-V_\beta^\star(h^o)=A_\beta^\star(h^o,a). \tag{68}$$

Using $A_\beta^\star(h^o,a)=\beta\log\frac{\pi_\beta^\star(a|h^o)}{\mu_{\mathrm{ref}}(a|h^o)}$, we obtain

$$r^\star(h^o,a)-\beta\log\frac{\pi(a\mid h^o)}{\mu_{\mathrm{ref}}(a\mid h^o)}+\gamma\mathbb{E}\left[V_\beta^\star(h^{o\prime})\mid h^o,a\right]-V_\beta^\star(h^o)=-\beta\log\frac{\pi(a\mid h^o)}{\pi_\beta^\star(a\mid h^o)}. \tag{69}$$

Taking expectations along trajectories generated by π, summing with discount factors, and using telescoping of the optimal value terms yields

$$\begin{aligned}V_\beta^\star(h_0^o)-V_\beta^\pi(h_0^o)&=\beta\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\log\frac{\pi(a_t\mid h_t^o)}{\pi_\beta^\star(a_t\mid h_t^o)}\,\middle|\,h_0^o\right]\\&=\beta\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t D_{\mathrm{KL}}\left(\pi(\cdot\mid h_t^o)\,\|\,\pi_\beta^\star(\cdot\mid h_t^o)\right)\,\middle|\,h_0^o\right].\end{aligned} \tag{70}$$

Define the normalized discounted history occupancy measure $d_\pi^{h_0^o}(B) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\Pr_\pi\left(h_t^o\in B\mid h_0^o\right)$ for every measurable set $B\subseteq\mathcal{H}^o$. Then

$$\begin{aligned}V_\beta^\star(h_0^o)-V_\beta^\pi(h_0^o)&=\frac{\beta}{1-\gamma}\mathbb{E}_{h^o\sim d_\pi^{h_0^o}}\left[D_{\mathrm{KL}}\left(\pi(\cdot\mid h^o)\,\|\,\pi_\beta^\star(\cdot\mid h^o)\right)\right]\\&=\frac{1}{1-\gamma}\mathcal{R}_{d_\pi^{h_0^o}}(f_\eta,q).\end{aligned} \tag{71}$$

Applying Theorem 4.3 with $\nu=d_\pi^{h_0^o}$ gives

$$V_\beta^\star(h_0^o)-V_\beta^\pi(h_0^o)=\frac{\mathcal{E}_{\mathrm{act}}(f_\eta,q;d_\pi^{h_0^o})+\mathcal{E}_{\mathrm{rep}}(f_\eta;d_\pi^{h_0^o})}{1-\gamma}. \tag{72}$$

Dropping the nonnegative actor term proves Eq. (17). □

F Proofs for Observation Anchoring under Learnable Representations

Proof of Proposition 4.5. We prove the three statements separately.

Latent-coordinate invariance. By the transformed encoder and actor defined in part (i) of Proposition 4.5,

$$\begin{aligned}\pi_\theta^\varphi\left(a\mid f_\eta^\varphi(h^o)\right)&=\pi_\theta\left(a\mid\varphi^{-1}\left(\varphi(f_\eta(h^o))\right)\right)\\&=\pi_\theta\left(a\mid f_\eta(h^o)\right).\end{aligned} \tag{73}$$

Since the reference policy is defined on the observable history and is unchanged by φ,

$$\begin{aligned}A_{\eta,\theta}^{\mathrm{OA},\varphi}(h^o,a)&=\beta\log\frac{\pi_\theta^\varphi(a\mid f_\eta^\varphi(h^o))}{\mu_{\mathrm{ref}}(a\mid h^o)}\\&=\beta\log\frac{\pi_\theta(a\mid f_\eta(h^o))}{\mu_{\mathrm{ref}}(a\mid h^o)}\\&=A_{\eta,\theta}^{\mathrm{OA}}(h^o,a).\end{aligned} \tag{74}$$

The trajectory score is a discounted sum of these advantages. Therefore, $S_{\eta,\theta}^{\mathrm{OA},\varphi}(\tau^o)=S_{\eta,\theta}^{\mathrm{OA}}(\tau^o)$, and any objective depending only on differences of these scores, including OA-LSA-DPO, is also invariant.

Observation-anchored encoder gradient. Differentiating Eq. (9) gives

$$\begin{aligned}\nabla_\eta A^{\mathrm{OA}}_{\eta,\theta}(h^o, a) &= \beta \nabla_\eta \log \pi_\theta\left(a \mid f_\eta(h^o)\right) - \beta \nabla_\eta \log \mu_{\mathrm{ref}}(a \mid h^o) \\ &= \beta \nabla_\eta \log \pi_\theta\left(a \mid f_\eta(h^o)\right), \end{aligned} \tag{75}$$

because μ_{ref} is independent of η. Applying the chain rule yields

$$\nabla_\eta A^{\mathrm{OA}}_{\eta,\theta}(h^o, a) = \beta J_\eta f_\eta(h^o)^\top \left.\nabla_z \log \pi_\theta(a \mid z)\right|_{z=f_\eta(h^o)}, \tag{76}$$

which proves Eq. (10).

To connect this result to the trajectory-level objective, let $\Delta S_{\eta,\theta} := S_{\eta,\theta}(\tau^{o,w}) - S_{\eta,\theta}(\tau^{o,l})$. For $\mathcal{L}_{\mathrm{pref}}(\eta, \theta) = -\mathbb{E}_{\mathcal{D}_{\mathrm{pref}}}\left[\log \sigma(\Delta S_{\eta,\theta})\right]$, direct differentiation gives

$$\nabla_\eta \mathcal{L}_{\mathrm{pref}} = -\mathbb{E}_{\mathcal{D}_{\mathrm{pref}}}\left[\sigma(-\Delta S_{\eta,\theta}) \nabla_\eta \Delta S_{\eta,\theta}\right]. \tag{77}$$

Furthermore,

$$\begin{aligned}\nabla_\eta S_{\eta,\theta}(\tau^o) &= \sum_{t=0}^{H-1} \gamma^t \nabla_\eta A^{\mathrm{OA}}_{\eta,\theta}(h^o_t, a_t) \\ &= \beta \sum_{t=0}^{H-1} \gamma^t J_\eta f_\eta(h^o_t)^\top \left.\nabla_z \log \pi_\theta(a_t \mid z)\right|_{z=f_\eta(h^o_t)}. \end{aligned} \tag{78}$$

Hence, a trajectory-level preference label can propagate through every latent state visited along the compared rollouts.

Co-encoded reference cancellation. Differentiating the co-encoded advantage defined in part (iii) of Proposition 4.5 gives

$$\begin{aligned}\nabla_\eta A^{\mathrm{co}}_{\eta,\theta,\bar\theta}(h^o, a) &= \beta \nabla_\eta \log \pi_\theta\left(a \mid f_\eta(h^o)\right) \\ &\quad - \beta \nabla_\eta \log \pi_{\bar\theta}\left(a \mid f_\eta(h^o)\right) \\ &= \beta J_\eta f_\eta(h^o)^\top \Big[\nabla_z \log \pi_\theta(a \mid z) - \nabla_z \log \pi_{\bar\theta}(a \mid z)\Big]_{z=f_\eta(h^o)}. \end{aligned} \tag{79}$$

At $\theta = \bar\theta$, the numerator and denominator of the co-encoded advantage are identical. Therefore, $A^{\mathrm{co}}_{\eta,\bar\theta,\bar\theta}(h^o, a) = \beta \log 1 = 0$. Their latent score gradients are also identical, so $\nabla_\eta A^{\mathrm{co}}_{\eta,\bar\theta,\bar\theta}(h^o, a) = 0$. Every corresponding trajectory score and its encoder gradient are therefore zero at reference initialization: $S^{\mathrm{co}}_{\eta,\bar\theta,\bar\theta}(\tau^o) = 0, \nabla_\eta S^{\mathrm{co}}_{\eta,\bar\theta,\bar\theta}(\tau^o) = 0$. Substitution into Eq. (77) shows that the preference-loss gradient with respect to the shared encoder also vanishes: $\nabla_\eta \mathcal{L}^{\mathrm{co}}_{\mathrm{pref}} = 0$ when $\theta = \bar\theta$. This cancellation does not occur structurally in Eq. (10), because the observation-history reference contributes no encoder derivative. □

Proof of Corollary 4.6. Let $c \in \{0, 1\}$ be uniformly distributed. By assumption, conditioned on any future action sequence, the physical future has the same distribution under $h^{o,0}$ and $h^{o,1}$. Hence,

$$P\left(o^o_{\mathrm{future}} \mid h^{o,0}, a_{\mathrm{future}}\right) = P\left(o^o_{\mathrm{future}} \mid h^{o,1}, a_{\mathrm{future}}\right). \tag{80}$$

A predictive representation can therefore map the two histories to the same latent state without losing information required to predict the physical future.

For the one-step decision reward $r_c(a) = \mathbf{1}\{a = c\}$ and uniform reference $\mu_{\mathrm{ref}}(a \mid h^{o,c}) = 1/2$, the optimal regularized policy is

$$\pi^\star_\beta(a \mid h^{o,c}) = \frac{\exp\left(\mathbf{1}\{a = c\}/\beta\right)}{1 + \exp(1/\beta)}. \tag{81}$$

Define $p := \frac{\exp(1/\beta)}{1+\exp(1/\beta)}$. The two optimal action distributions are $p_{h^{o,0}} = (p, 1-p)$ and $p_{h^{o,1}} = (1-p, p)$. Suppose the encoder discards the cue: $f_\eta(h^{o,0}) = f_\eta(h^{o,1}) = z$. By the geometric-barycenter result in Theorem 4.3, the optimal latent actor satisfies

$$q^\dagger_{\eta,\nu}(a \mid z) \propto \sqrt{p_{h^{o,0}}(a) p_{h^{o,1}}(a)}$$

$$= \sqrt{p(1-p)} \tag{82}$$

for both actions. Thus, $q^{\dagger}_{\eta,\nu}(\cdot \mid z) = \left(\frac{1}{2}, \frac{1}{2}\right)$. The associated geometric normalizer is

$$\begin{aligned} Z^{\nu}_{\eta}(z) &= 2\sqrt{p(1-p)} \\ &= \frac{2\exp(1/(2\beta))}{1+\exp(1/\beta)} \\ &= \frac{1}{\cosh(1/(2\beta))}. \end{aligned} \tag{83}$$

Therefore, the irreducible representation error is

$$\inf_{q} \mathcal{R}_{\nu}(f_{\eta}, q) = -\beta \log Z^{\nu}_{\eta}(z) = \beta \log \cosh\left(\frac{1}{2\beta}\right) > 0. \tag{84}$$

Finally, consider the one-bit representation $f(h^{o,c}) = c$. Each reachable latent state now corresponds to exactly one target action distribution. Therefore, $Z^{\nu}_{f}(c) = 1$ for $c \in \{0, 1\}$, and Theorem 4.3 gives $\mathcal{E}_{\mathrm{rep}}(f; \nu) = 0$. Hence, one bit of preference memory is sufficient to eliminate the otherwise irreducible control error. □

G Additional Experimental Evidence and Analyses

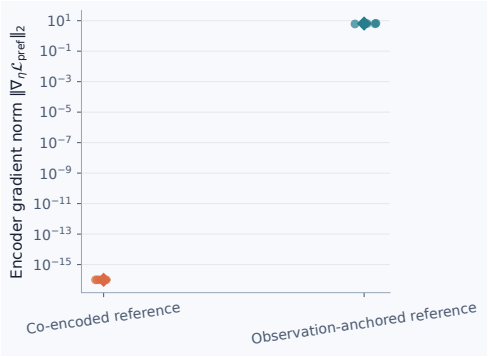
This appendix expands the main results into a mechanism-to-behavior evaluation. Controlled numerical tests first isolate the representation-level failure caused by history aliasing. MiniGrid then tests whether the learned state preserves a transient request under exact later-observation aliasing, while ManiSkill tests whether the same principle transfers to heterogeneous offline preferences and continuous robot control. We complement aggregate success with learning curves, preference-resolved outcomes, counterfactual reversals, latent geometry, time-resolved displacement, full 2D/3D trajectories, component ablations, and sensitivity analyses. This breadth is important: the central claim concerns the construction of a preference-sufficient state, which cannot be established by a single terminal score. All stochastic benchmark results are aggregated over five independent random seeds; uncertainty is reported as the standard error of the mean where shown.

G.1 Evaluation Scope and Comparisons

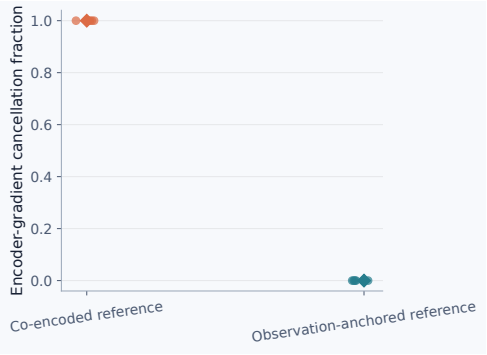
Baselines. We compare OA-LSA-DPO against three alternatives that separate representation adaptation from the choice of preference-learning objective. *Dyn-only LSA-DPO* applies the same trajectory-level preference objective while keeping the predictive representation fixed. *Frozen-latent RLHF* learns a scalar reward on the fixed latent state and optimizes a policy against that reward. *Adaptive-latent RLHF* updates the preference-facing state within the same reward-model pipeline. The resulting comparison separates direct trajectory preference optimization, representation adaptation, and explicit reward modeling. It therefore tests whether fitting preference labels is enough, or whether preference supervision must reshape the state consumed by the actor. Importantly, the observation-space anchor is a reference construction rather than an additional learned model: it keeps the comparison independent of the trainable encoder while reusing the same pretrained predictive backbone. Thus the proposed method changes where preference supervision enters the control pipeline without introducing a separately trained reward model.

Evaluation scope. Preference supervision is obtained from offline trajectory comparisons, and the final metrics are evaluated on held-out contexts and rollouts. We report both aggregate and preference-resolved behavior because task completion can remain high even when a policy ignores the requested outcome. This distinction is central to our evaluation: a method is successful only when the earlier request changes the later action and the corresponding physical outcome.

MiniGrid. A transient request selects one of two goals, but later observations at the decision point are identical. Success, balance, goal confusion, and a same-layout cue-swap test jointly determine whether the policy remembers the request and reverses the aliased decision for the correct reason.

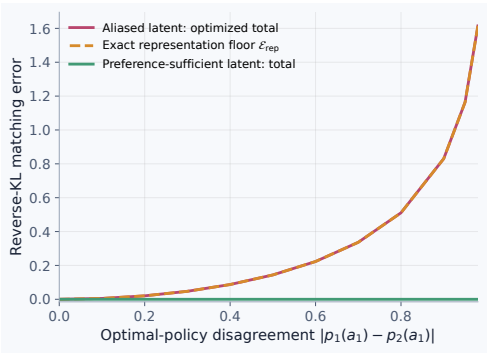


(a) Encoder gradient at the aliased initialization.

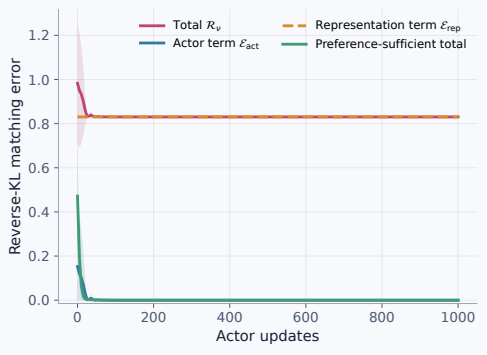


(b) Cancellation under symmetric supervision.

Figure 2: Numerical verification of the encoder-gradient mechanism. When two histories collapse to the same state while requiring opposite actions, their policy-level learning signals cancel at the representation. The result explains why additional actor optimization cannot by itself recover the missing preference variable.



(a) Aliasing floor versus policy disagreement.



(b) Actor optimization reaches the floor.

Figure 3: A fixed aliased state imposes a representation-dependent error floor. Increasing actor optimization reduces ordinary fitting error but cannot remove the ambiguity induced by a state that maps preference-incompatible histories to the same decision input. This motivates adapting the state rather than only strengthening the downstream optimizer.

ManiSkill. A transient request selects which of two cubes should be pushed to the goal under heterogeneous continuous-control trajectories. We evaluate target and balanced success, wrong-object behavior, preference reversal, latent organization, object displacement, terminal geometry, and complete robot trajectories. These complementary views test whether preference information becomes a persistent decision variable rather than merely a predictable label.

G.2 Controlled Numerical Verification of the Mechanism

The following controlled calculations instantiate the failure mode analyzed in the theory without environmental confounders. They are not additional task benchmarks; their purpose is to connect the optimization geometry directly to the behavioral experiments. The first diagnostic tests whether opposing preference branches cancel at an aliased state; the second tests whether additional actor optimization can overcome the resulting representation floor. Together they isolate the specific failure that motivates joint state and policy learning. These diagnostics analyze the learning signal and do not add an extra module to OA-LSA-DPO.

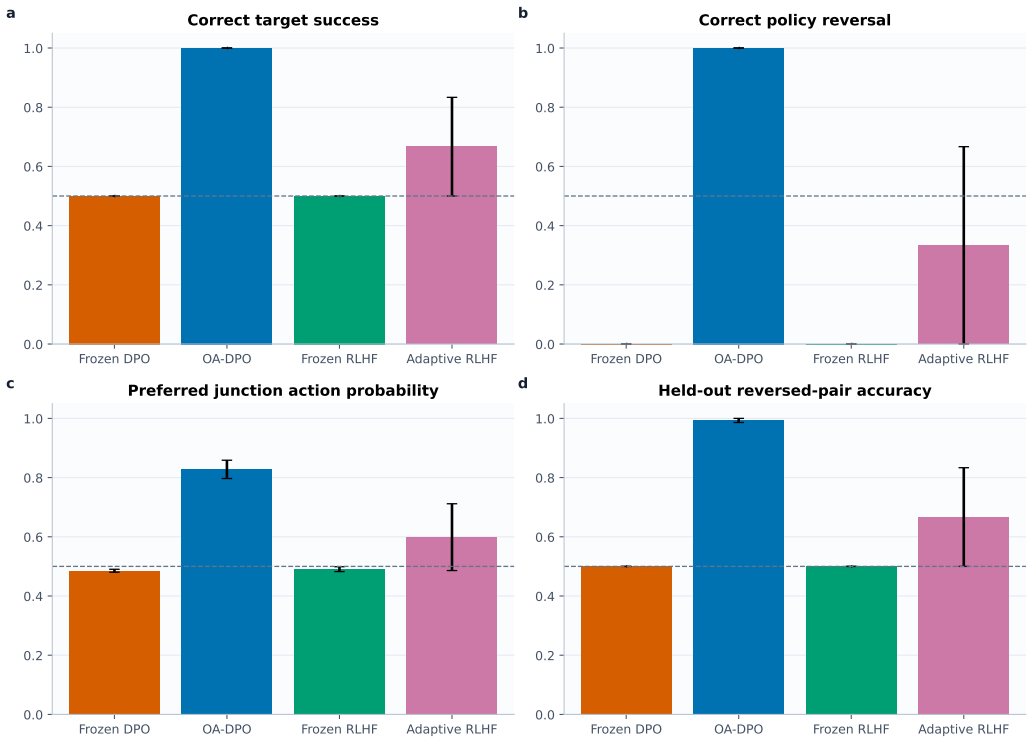


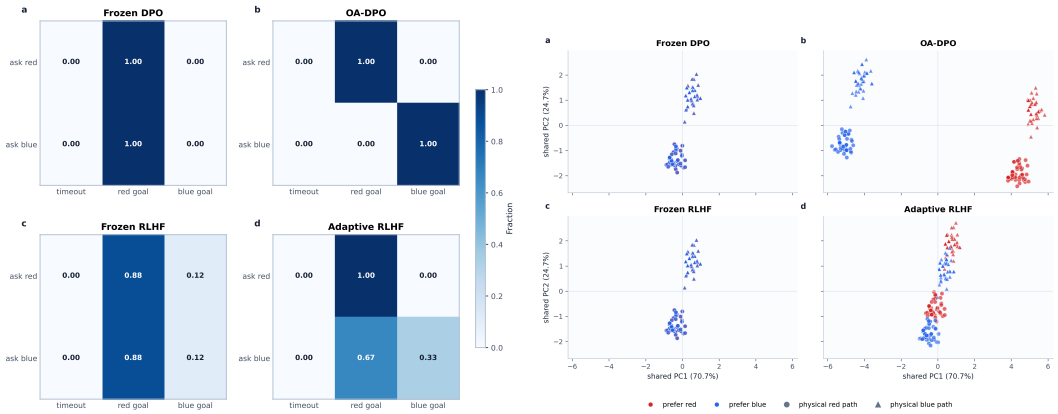
Figure 4: Primary MiniGrid learning and behavior results. Aggregate completion is shown together with preference-resolved success so that policies that solve the maze while ignoring the request are not credited as preference-grounded. OA-LSA-DPO couples optimization progress to balanced requested-goal behavior.

G.3 MiniGrid: Preference Memory under Exact Observation Aliasing

Learning dynamics and final behavior. The learning curves and preference-resolved outcomes show that task success is not sufficient evidence of preference grounding. Fixed-state methods can learn a reusable route yet continue to select the same branch under both requests. In contrast, OA-LSA-DPO improves requested-goal and balanced success together, which is the signature expected when preference supervision repairs the decision state rather than simply increasing generic task competence. The result operationalizes preference sufficiency behaviorally: two histories that lead to the same later observation remain distinguishable exactly when their optimal preference-conditioned decisions differ.

Failure diagnosis and latent organization. The goal-confusion matrix identifies the behavioral consequence of aliasing, while the latent projection tests the corresponding internal organization. Frozen representations preserve the symmetry and therefore reuse a decision. Adaptive reward learning introduces partial separation but does not reliably align it with the requested intervention. OA-LSA-DPO organizes the latent state by both remembered request and relevant trajectory outcome, matching the structure required by preference sufficiency. This paired analysis links an internal representational change to the corresponding reduction in behavioral confusion, rather than treating latent visualization as standalone evidence.

Counterfactual reversibility. The strongest test keeps the map, start state, and later observation sequence fixed while changing only the earlier request. A preference-sufficient state must cause the policy to reverse the branch selected at the aliased junction. Because every later physical input is held fixed, successful reversal rules out layout memorization and directly attributes the decision change to retained preference history.



(a) Requested goal versus reached goal. **(b)** Held-out latent organization.

Figure 5: Behavioral and representational diagnosis in MiniGrid. The two views connect the reduction in goal confusion to a learned state that preserves the otherwise transient preference information.

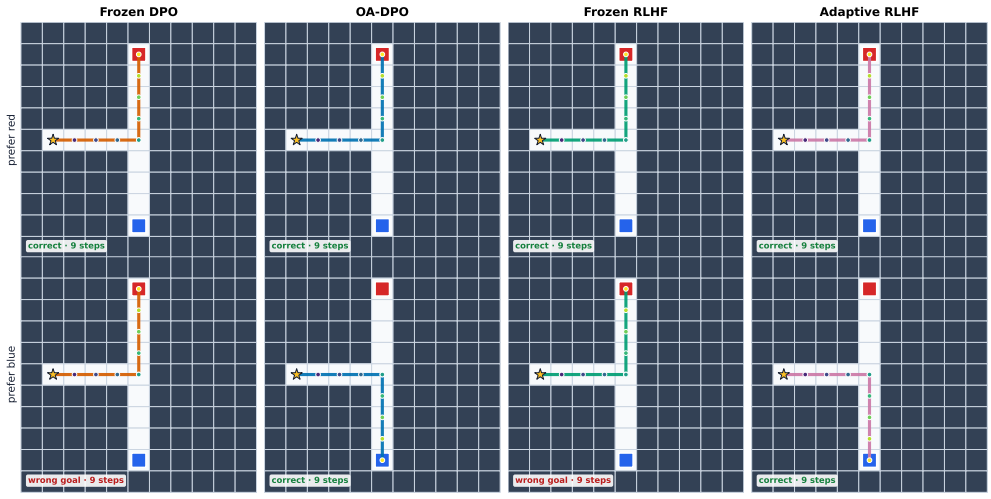


Figure 6: Same-layout counterfactual paths in MiniGrid. Only the transient request is changed. The resulting route reversal demonstrates that OA-LSA-DPO uses remembered preference information rather than a layout-specific action shortcut.

G.4 ManiSkill: From Preference Geometry to Physical Control

Preference structure in the learned state. We first inspect the common latent coordinates using only held-out rollouts. The desired organization is not arbitrary class separation: the state must distinguish the requested target while retaining the scene information needed for control. Frozen dynamics features remain dominated by trajectory geometry, whereas OA-LSA-DPO creates a preference-aligned partition that supports different actions in otherwise comparable scenes. Adaptive RLHF changes the geometry, but its clusters are less consistently tied to the requested outcome. This is the continuous-control counterpart of the MiniGrid aliasing test: preference information must become available in the actor-facing coordinates at the time an irreversible contact decision is made.

Persistence over the trajectory. A reset-time separation would be insufficient if it vanished before the robot's decisive interaction with the cubes. The time-resolved latent view therefore tracks preference organization throughout execution. Under OA-LSA-DPO, the relevant distinction persists after the cue is no longer present, linking the learned geometry to the later control decision. Persistence, not initial separability alone, is what distinguishes a useful preference state from a transient preference classifier.

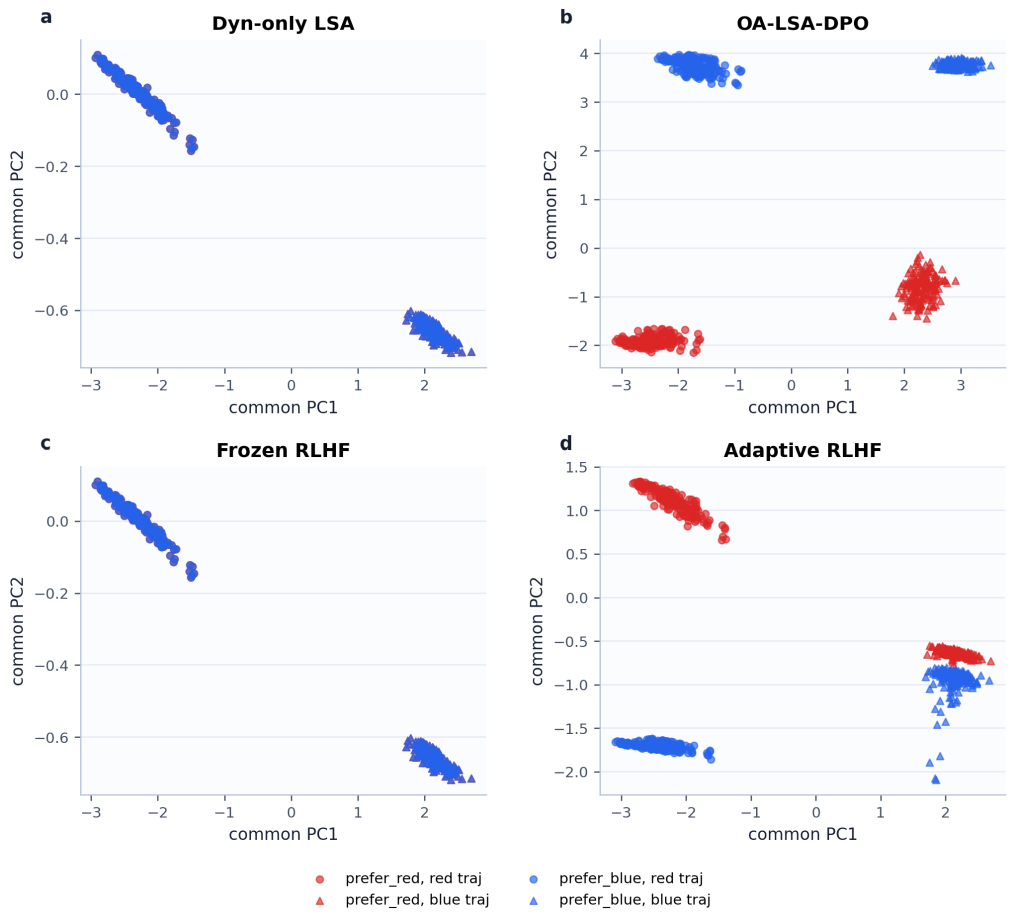


Figure 7: Held-out ManiSkill latent projections for the evaluated methods. Colors denote the requested cube and markers denote the realized trajectory type. OA-LSA-DPO forms preference-relevant structure that is absent from frozen representations and more behaviorally aligned than unconstrained end-to-end reward adaptation.

Terminal outcomes and displacement composition. We next test whether the learned distinction changes the physical outcome. Final cube-to-goal distance measures progress toward the requested goal, while the displacement decomposition reveals whether the non-requested cube was moved instead. This pairing is important because a policy may obtain partial task progress while acting on the wrong object. OA-LSA-DPO most consistently turns preference separation into selective requested-cube motion. The joint distance and displacement analysis therefore tests both efficacy and semantic correctness of the learned control behavior.

When the requested motion is produced. The displacement curves localize the difference in time. They distinguish a method that commits early to the requested object from one that accumulates incidental motion or corrects only after interacting with the wrong cube.

Trajectory-level qualitative evidence. Top-down and three-dimensional trajectories expose failure modes that terminal success alone can hide, including contact with the wrong cube, inefficient approaches, and late recovery. The views below use the same semantic color coding across methods and complement the aggregate metrics rather than replacing them. Taken together with the time-resolved displacement curves, they show how the learned preference state changes the complete interaction, not only its terminal label.

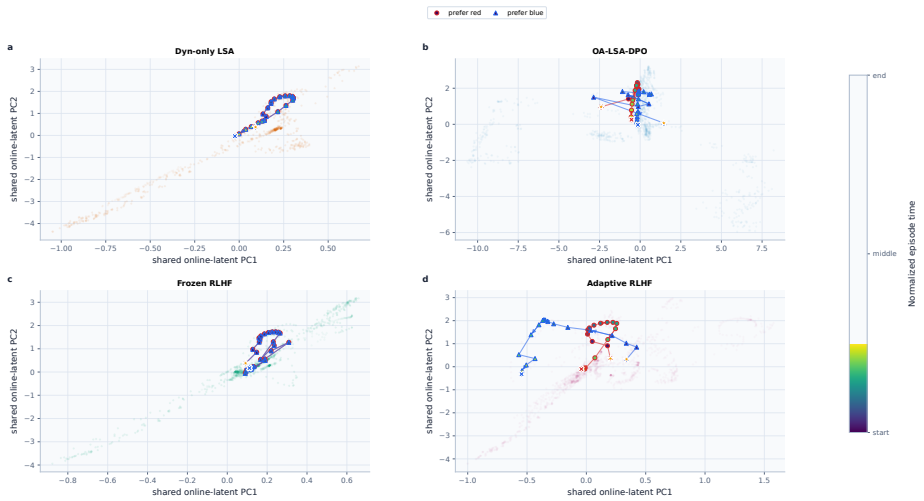
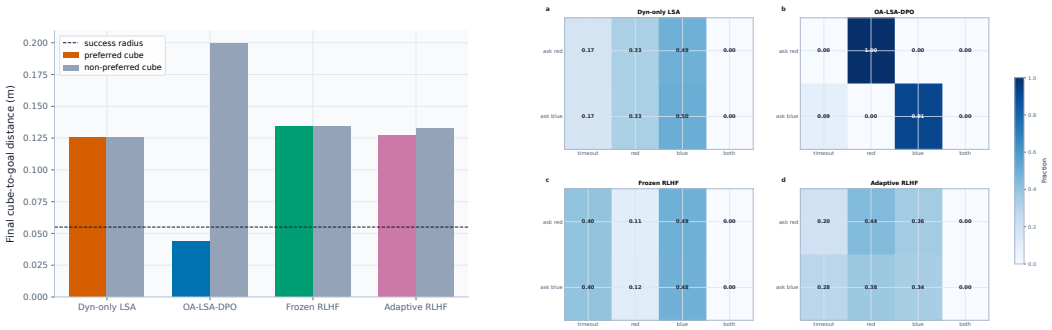


Figure 8: Time-resolved latent organization in ManiSkill. Preference-dependent structure remains available at the stages where the robot approaches and pushes the requested cube, rather than appearing only at initialization.



(a) Final requested-cube distances.

(b) Outcome and displacement composition.

Figure 9: Complementary physical-outcome diagnostics in ManiSkill. The left panel measures how close the requested cube finishes to its goal; the right panel separates requested motion from wrong-cube motion. Together they show that the gain of OA-LSA-DPO is preference-specific rather than generic scene disturbance.

Terminal robot–object geometry. The final three-dimensional states provide a direct physical sanity check on the aggregate success metrics. They show whether the robot finishes in a configuration consistent with selectively moving the requested cube, rather than merely producing a favorable scalar score.

G.5 Ablations: Why the Proposed Mechanism Works

The ablations below test two distinct questions. The trajectory-score study asks which temporal evidence must enter the preference objective. The architecture study asks where representation adaptation can be introduced without sacrificing the predictive structure inherited from dynamics pretraining. These two axes isolate the temporal and representational parts of the contribution: trajectory accumulation determines how preference evidence is assigned across the interaction, while the dedicated adapter determines how that evidence enters the state used by the actor.

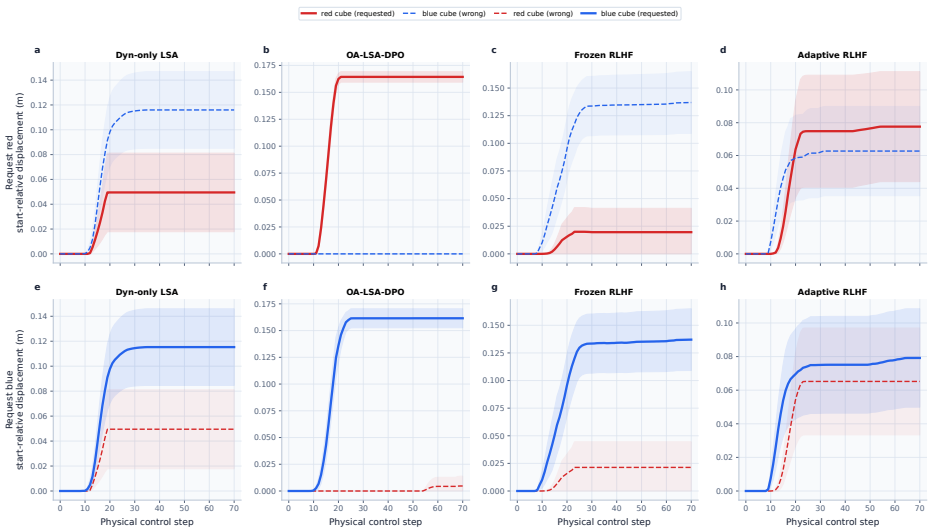


Figure 10: Requested- and non-requested-cube displacement over time. OA-LSA-DPO concentrates motion on the object selected by the transient request, connecting its persistent latent separation to the timing of the physical intervention.



Figure 11: Top-down ManiSkill trajectories. The planar view makes the selected object, approach direction, and final displacement directly comparable across methods and preference conditions.

G.5.1 TRAJECTORY-SCORE ABLATION

The contrast with the final-step variant is the important result: preference grounding depends on assigning credit across the interaction, not merely on classifying the terminal configuration. The comparable behavior of several accumulated scores shows that the advantage is not tied to a narrow hand-tuned temporal weighting, while the theory-derived default is numerically strongest across the primary preference-control metrics. We therefore obtain a principled default without sacrificing the robustness shared by alternative trajectory-accumulation rules.

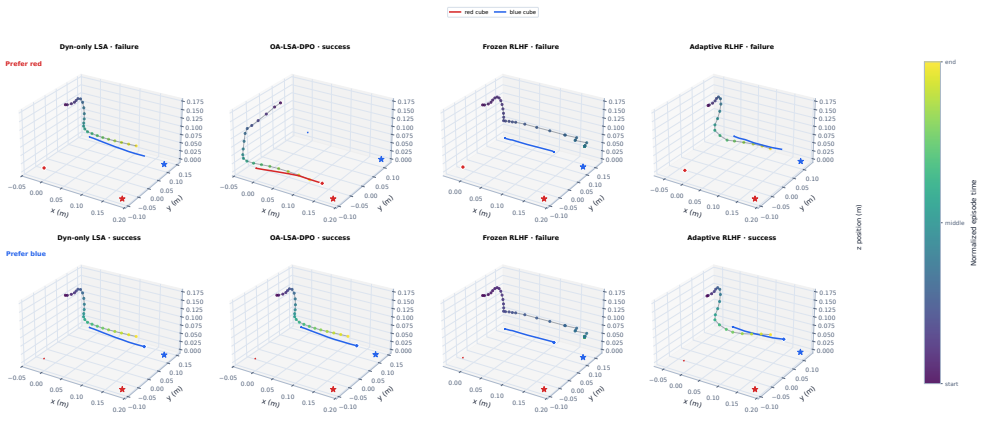


Figure 12: Three-dimensional ManiSkill trajectories. This view confirms that the preference-dependent differences seen in the planar plots correspond to distinct robot-object interactions in the full continuous-control scene.

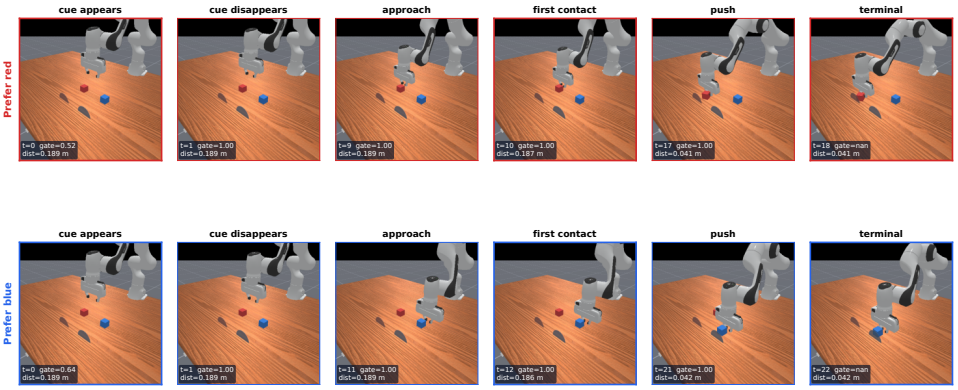


Figure 13: Representative ManiSkill rollout storyboard. The frame sequence shows how the reset-time request changes the later approach and contact choice after that request is no longer directly observable.

G.5.2 Representation-adaptation ablation

These results support a separation-of-roles interpretation. The dynamics encoder supplies stable predictive coordinates; a dedicated preference adapter adds the missing distinction needed by the actor; and the fixed observation-space anchor constrains policy change without inheriting drift from the adapted encoder. The adapter is therefore not simply a larger policy network: it changes which histories are distinguishable while leaving the pretrained predictive state intact. Because both adapter variants retain the same frozen predictive backbone, they share the same predictive-state diagnostics; the full default nevertheless achieves the strongest preference accuracy and target success. The shared-encoder variants expose the opposite tradeoff: preference fitting can improve while the predictive coordinates needed for reliable control are substantially distorted.

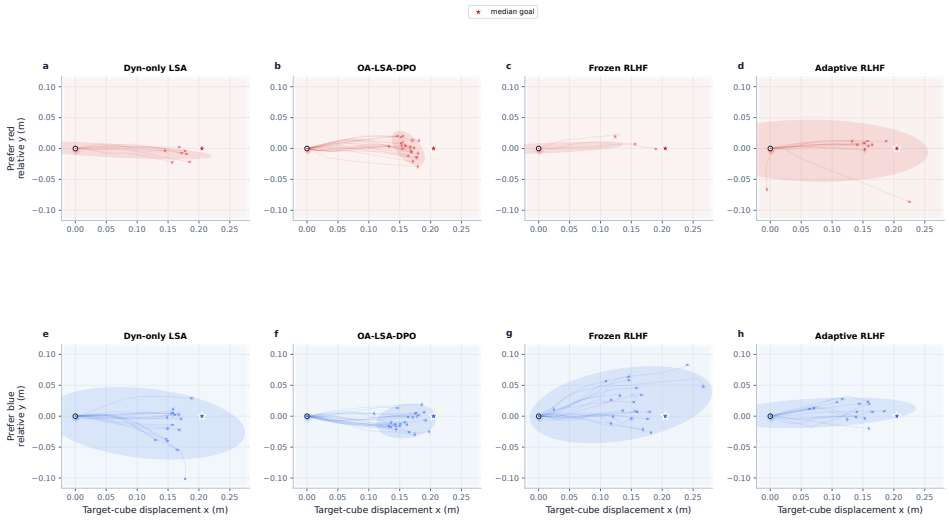


Figure 14: Cross-scene trajectory geometry. Consistent preference-dependent displacement across varied initial configurations argues against memorizing a single approach path or exploiting a fixed spatial shortcut.

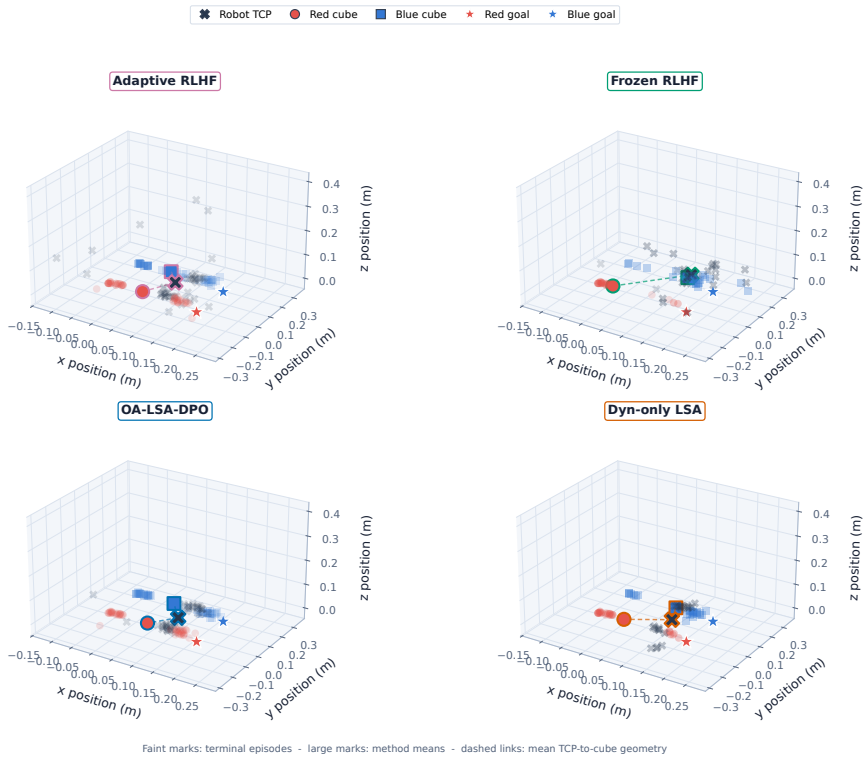
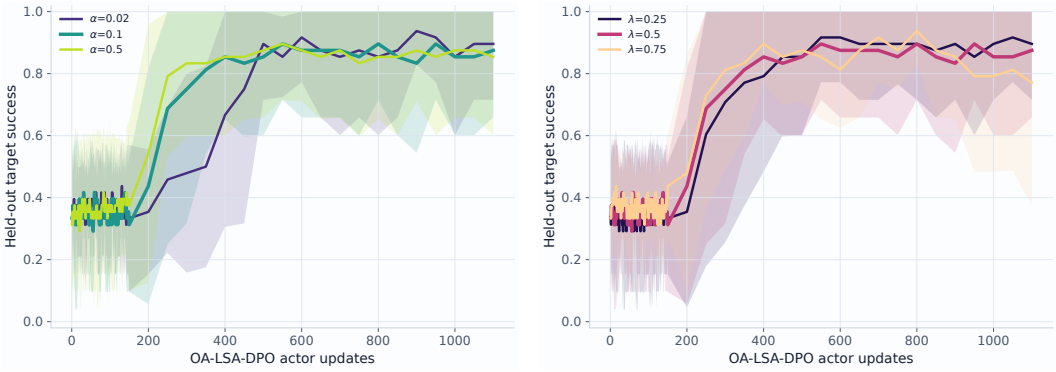


Figure 15: Final robot and cube states by method. The terminal geometry makes preference-specific object selection visible and complements the distance, success, and displacement summaries.

G.5.3 Why an Observation-Space Reference Is Essential

The main paper already reports the complete anchor ablation. Repeating an expanded table here would duplicate the same behavioral conclusion while placing disproportionate emphasis on raw



(a) Observation-anchor coefficient. **(b)** Representation-update coefficient.

Figure 16: ManiSkill sensitivity analysis. Performance remains stable across a nontrivial range around the default configuration, indicating that the main result follows from the proposed factorization rather than a single finely tuned scalar value.

Table 5: ManiSkill trajectory-score ablation over five independent random seeds; target success is reported as mean ± SEM. All temporally accumulated variants preserve strong preference-conditioned control, whereas using only the terminal score discards the sequential evidence needed to identify the requested intervention. The blue row is the default score used by OA-LSA-DPO because it follows the sequential preference objective directly, without selecting a task-specific heuristic.

	Preference-grounded control			
Trajectory score	**Preference accuracy ↑**	**Target success ↑**	**Balanced success ↑**	**Wrong-cube rate ↓**
Theory-derived (default)	0.865	0.962±0.034	0.922	0.000
Normalized accumulation	0.863	0.900±0.081	0.800	0.000
Early-weighted accumulation	0.857	0.958±0.037	0.917	0.000
Final step only	0.832	0.379±0.090	0.067	0.404

latent scale, which is not itself a control objective. The combined evidence from the main anchor ablation, Fig. 2, and Figs. 7–8 gives the more informative picture. A moving latent reference can fit preference comparisons while its notion of proximity changes with the encoder; freezing the state avoids drift but cannot repair aliasing; and the observation-space anchor permits preference-driven state adaptation while keeping the reference independent of that adaptation. Because the anchor is evaluated from observable history rather than through a second trainable representation, this stability does not require an additional reward model or a duplicated latent learning pipeline. It is the design choice that connects the paper's theoretical failure mode to its best behavioral result.

G.6 Sensitivity and Optimization Diagnostics

Sensitivity to objective weights. We vary the observation-anchor coefficient and the representation-update coefficient around the default setting. The goal is not to identify a single isolated optimum, but to verify that the preference-grounding effect persists over a useful neighborhood of hyperparameters.

G.7 Experimental Synthesis

The appendix evidence supports one connected conclusion. The controlled experiments show that actor optimization encounters a floor when incompatible histories share a state. MiniGrid then demonstrates exact counterfactual reversibility under a cue swap in an otherwise fixed layout. ManiSkill shows that the same mechanism scales to continuous interaction: preference structure persists through time, selects the requested object, and appears in both aggregate outcomes and full trajectories. Finally, the ablations localize the gain to sequential preference credit, representation adaptation, and an encoder-independent anchor. The resulting evidence spans controlled optimization, discrete counterfactual control, continuous manipulation, representation geometry, physical outcome decomposition, full trajectories, and robustness to design choices. Thus the contribution is not simply

Table 6: ManiSkill representation-adaptation ablation over five independent random seeds; target success is reported as mean ± SEM. The dedicated preference adapter obtains the strongest joint control and predictive-preservation profile. Freezing prevents preference information from entering the decision state, while unconstrained shared updates damage the dynamics geometry. Removing the auxiliary predictive term retains most of the control gain, showing that the improvement is driven primarily by the dedicated preference-adaptation pathway rather than by an additional prediction loss.

	Preference-grounded control			**Predictive-state preservation**		
Variant	**Preference accuracy ↑**	**Target success ↑**	**Latent separation ↑**	**Reconstruction error ↓**	**One-step error ↓**	**Multi-step error ↓**
Frozen encoder	0.627	0.446 ± 0.054	0.000	0.025	0.006	0.004
Dedicated preference adapter (default)	0.865	0.962 ± 0.034	9.665	0.025	0.006	0.004
Shared end-to-end encoder	0.809	0.854 ± 0.091	1.859	4.519	14.791	19.853
Shared encoder + predictive loss	0.827	0.858 ± 0.082	1.177	0.132	0.244	0.342
Adapter without predictive loss	0.815	0.946 ± 0.036	6.262	0.025	0.006	0.004

a stronger preference classifier or reward model; it is a control state whose learned distinctions are sufficient for the preference-dependent decisions the actor must make.