HAMILTONIAN TEMPORAL CONTRASTIVE CURIOSITY FOR REINFORCEMENT LEARNING EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient exploration is essential for reinforcement learning in physical control, where useful behavior depends on temporally extended dynamics. Existing curiosity methods typically reward prediction error, state novelty, uncertainty, representation change, or instantaneous physical variation. Hamiltonian-based exploration further introduces physical structure by rewarding changes in a learned Hamiltonian across individual transitions, but such signals remain local in time and do not directly reveal whether an action reaches an informative future over longer horizons. This raises a key question: how can exploration identify action-conditioned futures that remain poorly captured over multiple temporal scales while reflecting structured variation in the underlying dynamics? To address this challenge, we propose **Hamiltonian Temporal Contrastive Curiosity** (**HTC**^{2}), which formulates exploration as temporal future discovery in a learned Hamiltonian phase space. HTC^2 trains a horizon-conditioned contrastive critic on same-episode future pairs and uses the positive-pair representation distance to prioritize action-conditioned futures that remain insufficiently organized by the current temporal representation. Multi-step Hamiltonian variation further provides a bounded structural modulation of this temporal curiosity signal, allowing exploration to account for structured changes in the learned phase dynamics without treating energy variation itself as the exploration objective. We evaluate HTC^2 on six DeepMind Control Suite tasks with SAC and PPO against representative exploration methods including ICM, RIDE, E3B, and Plan2Explore, together with a single-step Hamiltonian-change curiosity variant, with additional coverage, ablation, and Hamiltonian diagnostics characterizing its exploration behavior and learned phase-space structure. A quick visual overview of this work is available at `https://iclr27-htc2.github.io/`.

1 INTRODUCTION

Exploration is a central challenge in reinforcement learning (RL), especially in physical control, where useful behaviors often emerge only through temporally extended interactions with the environment (Burda et al., 2018b; Pathak et al., 2017; Thrun, 1992; Stadie et al., 2015; Sutton et al., 1998; Mohamed et al., 2026). In such settings, an agent must discover not only unfamiliar states, but also how its actions shape future system dynamics (Toth et al., 2019; Garcia et al., 2025). For example, locomotion depends on coordinated changes in balance, contact, velocity, and momentum over multiple time steps rather than merely visiting visually different postures (Arzani et al., 2025; Noda et al., 2025; Toth et al., 2019; Garcia et al., 2025). Physical exploration therefore requires more than generic state-space coverage (Pathak et al., 2017; Mohamed et al., 2026; Burda et al., 2018b; Raileanu & Rocktäschel, 2020): the agent must identify action-conditioned futures that reveal informative dynamics.

Existing curiosity-driven methods promote exploration through prediction error, novelty, episodic uncertainty, or representation displacement (Pathak et al., 2017; Burda et al., 2018b; Raileanu & Rocktäschel, 2020; Henaff et al., 2022). These signals are often defined from local states or short transitions: prediction error can be affected by irrelevant stochasticity, novelty may encourage visitation without distinguishing dynamical relevance, and representation displacement does not necessarily separate coherent physical evolution from arbitrary feature change. Temporal contrastive learning provides a way to capture longer-range relations by learning representations between current state–action pairs and their future states (Ziarko et al., 2025; Hansen et al., 2022). Such relations

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

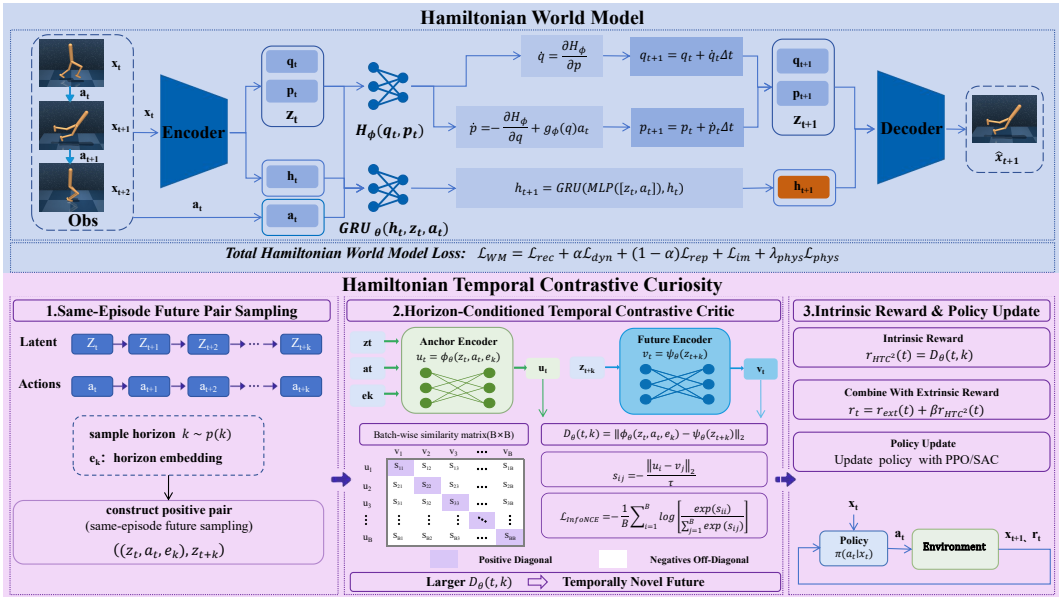


Figure 1: Overview of **Hamiltonian Temporal Contrastive Curiosity** (HTC^2). Observations are encoded into Hamiltonian phase states; same-episode future pairs train a horizon-conditioned temporal contrastive critic; and the resulting temporal discovery signal is modulated by bounded multi-step Hamiltonian variation.

can identify futures that remain poorly captured across multiple temporal scales, but they do not by themselves encode structured variation in the underlying physical dynamics.

Physical inductive biases provide a natural way to structure this temporal information. Hamiltonian world models represent dynamics through coordinate-like and momentum-like latent variables and organize their evolution with a learned Hamiltonian (Greydanus et al., 2019; Toth et al., 2019; Tang et al., 2026). Such phase-space representations describe how actions alter latent dynamics over time. DreamSAC exploits this structure by rewarding the change in the learned Hamiltonian across a single transition (Tang et al., 2026). However, single-step Hamiltonian-change curiosity remains local in time: a large instantaneous variation does not necessarily indicate that an action reaches an informative future over longer horizons. This motivates exploration that identifies informative action-conditioned futures across multiple temporal scales while accounting for structured variation in the learned dynamics.

To address this challenge, we introduce **Hamiltonian Temporal Contrastive Curiosity** (**HTC^2**), which formulates exploration as multi-step future discovery in a learned Hamiltonian phase space. As illustrated in Figure 1, HTC^2 trains a horizon-conditioned temporal contrastive critic on same-episode future pairs, using positive-pair representation distance as the primary curiosity signal. Bounded multi-step Hamiltonian variation then modulates this signal according to structured changes in the learned phase dynamics, while temporal novelty remains the exploration objective. A latent RND (Burda et al., 2018b) warm start stabilizes the initial learning stage before transitioning to the HTC^2 signal.

Our contributions are summarized as follows:

1. We formulate physical exploration as multi-step action-conditioned future discovery in a learned Hamiltonian phase space, moving beyond curiosity based only on local novelty or single-step Hamiltonian change.
2. We introduce HTC^2, using horizon-conditioned temporal contrastive distance as the primary curiosity signal and bounded multi-step Hamiltonian variation as structural modulation of learned phase dynamics.
3. We provide comprehensive empirical validation on DeepMind Control Suite tasks with SAC and PPO against ICM, RIDE, E3B, Plan2Explore, and a single-step Hamiltonian-change variant, covering control performance, physical-state coverage, parametric OOD robustness, ablations, and Hamiltonian diagnostics.

2 Related Work

Exploration in Reinforcement Learning. Exploration remains a central challenge in reinforcement learning, especially under sparse or delayed rewards (Sutton et al., 1998; Thrun, 1992; Mohamed et al., 2026). Intrinsic-motivation methods address this problem by rewarding novelty, uncertainty, or prediction difficulty (Houthooft et al., 2016; Ostrovski et al., 2017; Burda et al., 2018b; Pathak et al., 2017; Burda et al., 2018a; Lee et al., 2019; Guo et al., 2022). Count-based approaches favor rarely visited states (Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017), but scale poorly to high-dimensional continuous domains. Prediction-based methods instead use learned errors or uncertainty: ICM relies on forward prediction error (Pathak et al., 2017), RND measures error against a fixed random target (Burda et al., 2018b), and related methods exploit representation change, model disagreement, information gain, or episodic uncertainty (Raileanu & Rocktäschel, 2020; Pathak et al., 2019; Houthooft et al., 2016; Henaff et al., 2022). Such local novelty signals can be affected by irrelevant stochasticity, as illustrated by the "noisy-TV" problem (Pathak et al., 2019; Mavor-Parker et al., 2022). More recently, model-based exploration has leveraged learned world models to plan toward informative future states and improve sample efficiency (Sekar et al., 2020). However, such models are primarily learned as predictive or generative dynamics models and do not explicitly impose Hamiltonian structure or related physics-informed priors on the latent dynamics.

World Models in Reinforcement Learning. World models learn predictive representations of environment dynamics and are widely used in model-based reinforcement learning (Ha & Schmidhuber, 2018b; Hafner et al., 2020; Lyu et al., 2025; Hafner et al., 2023; 2025; Xu et al., 2022; Liu et al., 2023; Ha & Schmidhuber, 2018a; Hafner et al., 2019b; Zhang et al., 2026). Latent world models compress high-dimensional observations and predict future latent states for planning, imagination, or policy learning (Hafner et al., 2019a; Assran et al., 2025; 2023). The Dreamer family uses recurrent latent dynamics with predictive reconstruction objectives (Hafner et al., 2020; 2023; 2025), whereas TD-MPC-style methods emphasize task-relevant latent prediction for efficient control (Hansen et al., 2024; 2022; Janner et al., 2019). Learned dynamics are also used for intrinsic exploration through prediction error, inverse dynamics, and uncertainty (Pathak et al., 2017; 2019; Raileanu & Rocktäschel, 2020; Henaff et al., 2022). However, standard world models mainly capture statistical regularities and need not explicitly encode physical structure (Hafner et al., 2020; 2023; 2025; 2019a; Assran et al., 2025; 2023), nor does predictive modeling alone specify which future interactions are useful for exploration (Sekar et al., 2020; Tang et al., 2026). HTC^2 uses a Hamiltonian phase representation as the space for measuring temporal future relations, turning the world model into a structured substrate for intrinsic exploration.

Hamiltonian Dynamics for World Models. Physics-structured world models introduce inductive biases from classical mechanics to improve dynamics learning and generalization (Greydanus et al., 2019; Cranmer et al., 2020; Zhong et al., 2019; Toth et al., 2019; Arzani et al., 2025; Troch et al., 2025). Hamiltonian neural networks and symplectic models represent dynamics through coordinate-like and momentum-like variables governed by gradients of a learned Hamiltonian (Toth et al., 2019; Greydanus et al., 2019; Zhong et al., 2019). Recent work extends these priors to reinforcement learning and pixel-based world models (Schiewer et al., 2024; Deng et al., 2025; Tang et al., 2026). Their use for exploration, however, remains less developed; directly rewarding learned energy change can favor local perturbations rather than temporally extended discovery (Tang et al., 2026). HTC^2 instead uses Hamiltonian variation only as a bounded modulation, while the primary exploration signal is the temporal contrastive distance between current phase state–action pairs and their multi-step futures.

3 Method

We propose **Hamiltonian Temporal Contrastive Curiosity** (**HTC**^{2}), an intrinsic exploration method for reinforcement learning with Hamiltonian latent world models. HTC^2 encourages the agent to discover future phase states that remain difficult to organize under the current temporal representation, while using learned Hamiltonian variation to provide a bounded physics-guided modulation. As illustrated in Figure 1, HTC^2 consists of a Hamiltonian phase-space world model, a horizon-conditioned temporal contrastive critic, and a temporal discovery reward with latent RND warm start.

3.1 PRELIMINARIES

Controlled Hamiltonian Dynamics. We consider latent dynamics with a phase state $z_t = (q_t, p_t)$, where q_t and p_t are learned coordinate-like and momentum-like latent variables (Toth et al., 2019). A controlled Hamiltonian formulation separates the internal latent dynamics from the effect of external actions (Zhong et al., 2019). Given a learned Hamiltonian $H_\varphi(q, p)$ and a learned control input mapping $g_\varphi(q)$, the continuous-time dynamics are

$$\frac{dq}{dt} = \frac{\partial H_\varphi(q,p)}{\partial p}, \qquad \frac{dp}{dt} = -\frac{\partial H_\varphi(q,p)}{\partial q} + g_\varphi(q)a_t. \tag{1}$$

The first component describes the internal Hamiltonian flow, while $g_\varphi(q)a_t$ represents the action-dependent control contribution to the momentum-like dynamics.

For an autonomous Hamiltonian system without external control, $a_t = 0$, the Hamiltonian is constant along the continuous trajectory. With the control term in Equation 1, its time derivative becomes

$$\frac{dH_\varphi}{dt} = \nabla_p H_\varphi(q,p)^\top g_\varphi(q)a_t. \tag{2}$$

Thus, Hamiltonian variation separates the effect of external control from the internally conservative Hamiltonian flow (Toth et al., 2019). HTC^2 does not directly maximize this variation; instead, it uses the corresponding multi-step quantity as a bounded modulation of temporal future discovery.

Finite-Horizon Future Distribution. Temporal contrastive learning requires a distribution over future phase states reached from the current state–action pair (Ho & Ermon, 2016; Torabi et al., 2018; Eysenbach et al., 2020). Let $\mathcal{K} = \{1, \ldots, K\}$ denote the set of admissible future horizons. We define the finite-horizon future distribution as

$$p_\mathcal{K}(z_f \mid z_t, a_t) = \sum_{k=1}^{K} \omega_k \, p_\mathcal{T}\left(z_{t+k} = z_f \mid z_t, a_t\right), \tag{3}$$

where $p_\mathcal{T}$ denotes the trajectory distribution and ω_k is the probability of sampling horizon k.

HTC^2 uses a truncated geometric horizon distribution,

$$\omega_k = \frac{(1-\gamma_\mathrm{T})\gamma_\mathrm{T}^{k-1}}{1-\gamma_\mathrm{T}^{K}}, \qquad k \in \mathcal{K}, \tag{4}$$

which satisfies $\sum_{k=1}^{K} \omega_k = 1$. The parameter $\gamma_\mathrm{T} \in (0, 1)$ controls preference for nearer futures while retaining explicitly multi-step positive pairs. Each pair is constrained to lie within the same episode.

Hamiltonian Phase-Space World Model. HTC^2 uses a recurrent Hamiltonian latent world model to construct the phase representation on which temporal discovery is performed. Given an observation x_t and recurrent deterministic state h_t, the posterior model infers

$$q_\varphi(z_t \mid h_t, x_t), \qquad z_t = (q_t, p_t). \tag{5}$$

The stochastic phase state is divided into coordinate-like and momentum-like components. In implementation, both components are represented by categorical stochastic variables; their posterior expectations are used whenever continuous phase values are required by the Hamiltonian transition or the temporal critic.

The deterministic state is updated recurrently from the previous phase state and action. The controlled Hamiltonian dynamics in Equation 1 then provide a physics-structured phase prediction, which is used together with the recurrent context to parameterize the next stochastic prior:

$$p_\varphi\left(z_{t+1} \mid h_{t+1}, z_t, a_t\right). \tag{6}$$

The Hamiltonian transition is integrated using a leapfrog step, while a learned residual prior retains sufficient flexibility for high-dimensional visual control. Detailed transition equations are given in Appendix B.

The world model is optimized with

$$\mathcal{L}_\mathrm{WM} = \mathcal{L}_\mathrm{rec} + \alpha\mathcal{L}_\mathrm{dyn} + (1-\alpha)\mathcal{L}_\mathrm{rep} + \mathcal{L}_\mathrm{inv} + \lambda_\mathrm{phys}\mathcal{L}_\mathrm{phys}. \tag{7}$$

Here, $\mathcal{L}_{\text{rec}}$ grounds the posterior latent state in the observation space; $\mathcal{L}_{\text{dyn}}$ and $\mathcal{L}_{\text{rep}}$ provide balanced prior–posterior regularization; $\mathcal{L}_{\text{inv}}$ preserves action-relevant information; and $\mathcal{L}_{\text{phys}}$ aligns the Hamiltonian prediction with the posterior phase transition.

For HTC^2, the world model provides posterior phase trajectories $(z_t, z_{t+1}, \ldots)$ for temporal pair construction and the learned Hamiltonian $H_\varphi(z_t)$ for evaluating multi-step Hamiltonian variation. Temporal positives are formed from actually reached same-episode posterior futures.

3.2 Hamiltonian Temporal Contrastive Curiosity

HTC^2 uses horizon-conditioned temporal contrastive mismatch as its primary curiosity signal and bounded multi-step Hamiltonian variation to modulate this signal.

Horizon-Conditioned Temporal Contrastive Critic. Contrastive representation learning brings positive pairs closer while separating them from negative samples (Chopra et al., 2005; Oord et al., 2018; Wang & Isola, 2020; Saunshi et al., 2019; Hassani & Khasahmadi, 2020). In HTC^2, the anchor is (z_t, a_t, k), the reached same-episode future z_{t+k} forms the positive pair, and other minibatch futures serve as negatives. The horizon $k \sim \omega_k$ is sampled according to Equation 4, enabling the critic to learn future relations over multiple temporal scales.

The critic consists of an anchor encoder and a future encoder:

$$u_{t,k} = \phi_\theta(z_t, a_t, e_k), \qquad v_{t+k} = \psi_\theta(z_{t+k}). \tag{8}$$

where e_k is a learnable horizon embedding.

The temporal score is defined by the negative representation distance:

$$C_\theta((z_t, a_t, k), z_f) = -\left\|\phi_\theta(z_t, a_t, e_k) - \psi_\theta(z_f)\right\|_2. \tag{9}$$

Given a minibatch of B same-episode positive pairs, we optimize the temporal critic with the InfoNCE objective (Oord et al., 2018):

$$\mathcal{L}_{\text{TC}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(C_\theta((z^i_{t_i}, a^i_{t_i}, k_i), z^i_{t_i+k_i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(C_\theta((z^i_{t_i}, a^i_{t_i}, k_i), z^j_{t_j+k_j})/\tau\right)}. \tag{10}$$

The matched future forms the positive pair, whereas the remaining minibatch futures act as negatives. The temperature τ controls the sharpness of the contrastive classification.

The policy does not directly optimize $\mathcal{L}_{\text{TC}}$. Instead, HTC^2 extracts the positive-pair distance

$$D_\theta(t, k) = \left\|\phi_\theta(z_t, a_t, e_k) - \psi_\theta(z_{t+k})\right\|_2. \tag{11}$$

A large $D_\theta(t, k)$ indicates that the reached future remains poorly aligned under the current temporal representation. As these transitions are revisited, the temporal critic learns to better organize them through Equation 10, making the temporal mismatch an adaptive exploration signal.

Hamiltonian-Modulated Temporal Curiosity. For the same sampled future horizon k, HTC^2 evaluates the multi-step Hamiltonian variation

$$\Delta H_{t,k} = \left|H_\varphi(z_{t+k}) - H_\varphi(z_t)\right|. \tag{12}$$

Rather than directly maximizing $\Delta H_{t,k}$, HTC^2 converts it into the bounded modulation

$$G_H(t, k) = 1 + \alpha_H \tanh\left(\text{clip}\left(\frac{\Delta H_{t,k}}{\sigma_H + \epsilon}, 0, c_H\right)\right), \tag{13}$$

where σ_H denotes a running scale estimate, ϵ provides numerical stability, α_H controls the modulation strength, and c_H is the clipping threshold. Since the baseline value of the gate is one and its amplification is bounded, Hamiltonian variation reweights rather than replaces temporal discovery.

The raw HTC^2 reward is

$$r^{\text{raw}}_{\text{HTC}^2}(t) = D_\theta(t, k) G_H(t, k) - \lambda_{\text{sm}} \left\|a_t - a_{t-1}\right\|_2^2, \tag{14}$$

where λ_{sm} controls an action-smoothness penalty. The reward is normalized before being combined with the cold-start signal.

Early in training, both the phase representation and the temporal critic are still being learned. We therefore use normalized latent RND, $\bar{r}_t^{\mathrm{RND}}$, as a warm-start exploration signal. Let n denote the global environment step. The transition weight is

$$w_n = \mathrm{clip}\left(\frac{n - T_{\mathrm{warm}}}{T_{\mathrm{anneal}}}, 0, 1\right). \tag{15}$$

The final intrinsic reward is

$$r_t^{\mathrm{int}} = (1 - w_n)\bar{r}_t^{\mathrm{RND}} + w_n \bar{r}_t^{\mathrm{HTC}^2}, \tag{16}$$

where $\bar{r}_t^{\mathrm{HTC}^2}$ denotes the normalized HTC^2 signal. The reward used by the policy optimizer is therefore

$$\tilde{r}_t = r_t^{\mathrm{ext}} + \beta r_t^{\mathrm{int}}, \tag{17}$$

where β is the global intrinsic-reward scale. Optimizer-specific training procedures for SAC and PPO are provided in Appendix B.6.

4 Experiments

We evaluate HTC^2 on the DeepMind Control Suite (DMC) (Tunyasuvunakool et al., 2020) through five complementary studies. First, we compare control performance under SAC and PPO to assess whether HTC^2 improves exploration across policy optimizers. Second, we measure physical-state coverage to examine how broadly each method explores the underlying state space. Third, we evaluate zero-shot retention under parametric dynamics shifts to test policy robustness. Fourth, component ablations isolate the contribution of the main HTC^2 design choices. Finally, Hamiltonian diagnostics examine whether the learned phase representation exhibits the intended energy and latent structure. Additional results and diagnostics are reported in Appendix C.

4.1 Experimental Setup

Benchmarks, baselines, and evaluation. We evaluate HTC^2 on six tasks: `walker_walk`, `walker_run`, `walker_stand`, `hopper_hop`, `hopper_stand`, and `cheetah_run`. HTC^2 is evaluated with both SAC (Haarnoja et al., 2018) and PPO (Schulman et al., 2017), and compared with four representative exploration methods: ICM (Pathak et al., 2017), RIDE (Raileanu & Rocktäschel, 2020), E3B (Henaff et al., 2022), and Plan2Explore (Sekar et al., 2020). Within each policy optimizer, all methods use the same training budget and evaluation settings. Results are reported as mean $\pm$ standard deviation, with the number of random seeds specified for each experiment. Detailed numerical results and complete learning curves are provided in Appendix C.1.

4.2 Control Performance on DMC

Figure 2 shows that HTC^2 maintains strong control performance under both SAC and PPO in comparison with ICM, RIDE, E3B, and Plan2Explore. Across the twelve tasks, HTC^2 achieves the highest mean final return in eleven settings. Under SAC, HTC^2 ranks first on five of the six tasks and performs nearly identically to the strongest baseline on `walker_stand`; under PPO, it achieves the highest mean return on all six tasks. The consistency across the two optimization regimes indicates that the performance of HTC^2 is not specific to either off-policy replay or on-policy updates.

The gains are particularly visible on tasks requiring coordinated multi-step transitions, such as `walker_walk` and `hopper_hop`. This trend is consistent with the design of HTC^2: exploration is driven by the mismatch between the current state–action representation and its reached multi-step future, while Hamiltonian variation only provides a bounded modulation of this temporal signal. Detailed per-task returns and complete SAC/PPO learning curves are reported in Appendix C.1.

4.3 Physical-State Exploration Coverage

We evaluate physical-state coverage on three DMC tasks using low-dimensional privileged simulator states, which are used only for evaluation and are not provided to the policy or intrinsic-reward module. For environment $\mathcal{E}$, let s_t^{phy} denote the task-specific physical state and $b_{\mathcal{E}}(\cdot)$ its fixed grid discretization. The cumulative coverage at training step t is

$$\mathcal{C}_t = \left|\left\{b_{\mathcal{E}}(s_\tau^{\mathrm{phy}}) \;:\; 0 \le \tau \le t\right\}\right|, \tag{18}$$

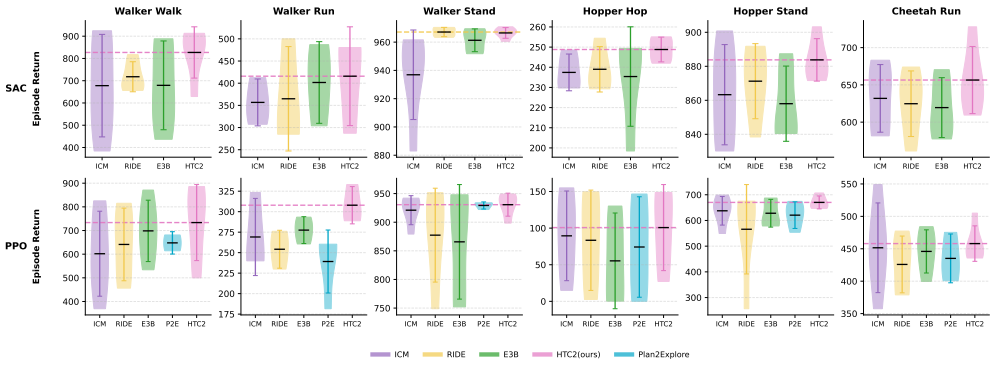


Figure 2: **Control performance of HTC^2 on six DMC tasks with SAC and PPO.** The figure summarizes episode-return performance across five random seeds for ICM, RIDE, E3B, Plan2Explore, and HTC^2. Numerical results are reported as mean $\pm$ standard deviation in Appendix C.1.

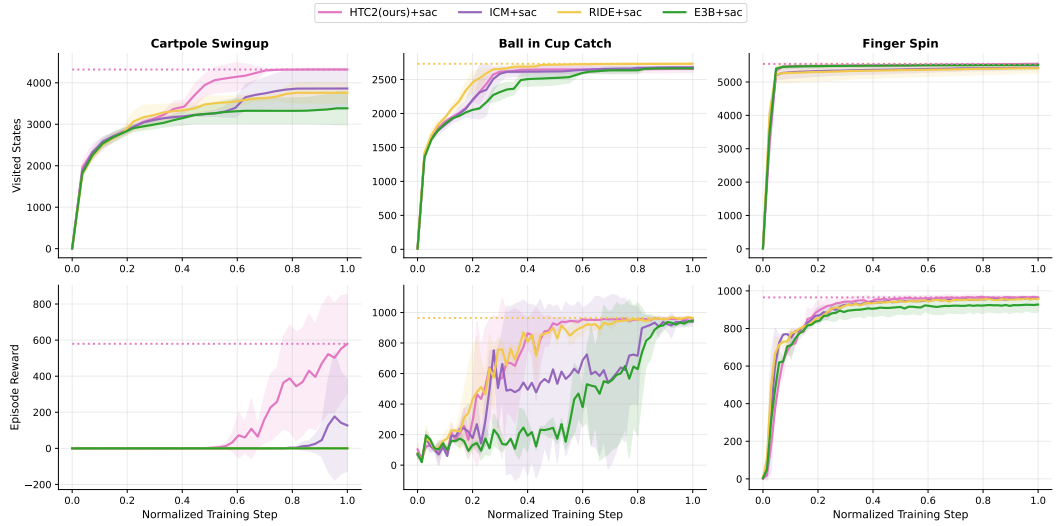


Figure 3: **Physical-state coverage and episode reward during training.** Top: cumulative unique discretized states; bottom: episode reward on `cartpole_swingup`, `ball_in_cup_catch`, and `finger_spin`. Curves show mean $\pm$ one standard deviation over four random seeds.

i.e., the number of unique discretized physical states visited.

For `cartpole_swingup`, we use the cart position and wrapped pole angle, $(x_{\text{cart}}, \theta_{\text{pole}})$. For `ball_in_cup_catch`, we use the ball position relative to the cup, $(x_{\text{ball}} - x_{\text{cup}}, z_{\text{ball}} - z_{\text{cup}})$. For `finger_spin`, we use the wrapped object angle and clipped angular velocity, $(\theta_{\text{obj}}, \dot{\theta}_{\text{obj}})$. Discretization details are provided in Appendix C.2.

Figure 3 shows distinct coverage patterns across the three environments. On `cartpole_swingup`, HTC^2 continues expanding its visited-state set after the other methods begin to saturate, reaching the highest final coverage and mean episode reward. On `ball_in_cup_catch`, RIDE attains the highest coverage, while HTC^2 ranks second in coverage AUC, final reward, and reward AUC. On `finger_spin`, HTC^2 achieves the highest final coverage and final reward, with a coverage AUC close to E3B.

Overall, HTC^2 exhibits strong physical-state coverage on `cartpole_swingup` and `finger_spin`, while remaining competitive on `ball_in_cup_catch`. Full coverage and reward statistics are reported in Appendix C.2.

4.4 Hamiltonian Mechanism Validation

Beyond policy learning and exploration coverage, we examine whether the learned Hamiltonian phase representation exhibits the structure used by HTC^2. Specifically, we test whether $H_\varphi(q, p)$

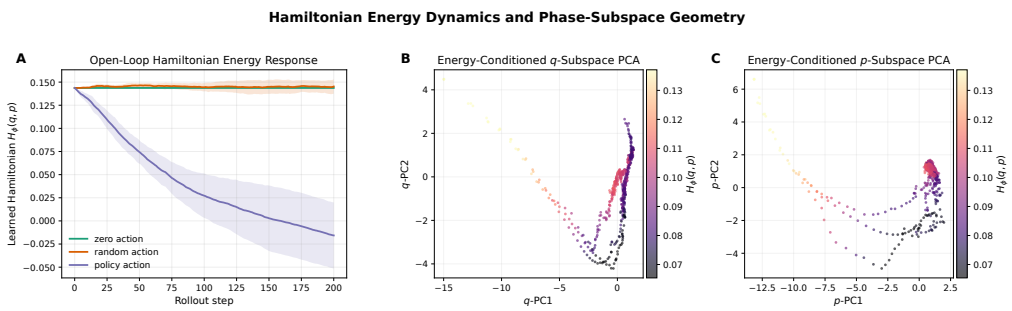


Figure 4: **Hamiltonian energy dynamics and phase-subspace geometry. A**: open-loop learned Hamiltonian evolution under zero, random, and policy actions. **B–C**: PCA projections of the coordinate-like latent q and momentum-like latent p, colored by $H_\varphi(q,p)$.

remains stable under passive evolution, responds to policy-induced control, and varies coherently over the coordinate-like and momentum-like latent subspaces.

(A) Controlled Hamiltonian energy response. Figure 4A compares the learned Hamiltonian under three action regimes. Under zero action, $H_\varphi(q,p)$ remains nearly flat over the rollout horizon, consistent with approximate energy preservation under autonomous Hamiltonian evolution. Random actions produce similarly bounded deviations around the zero-action trajectory. In contrast, policy actions induce a clear and sustained energy displacement with greater long-horizon variability, showing that the learned Hamiltonian responds differently to structured control.

(B) Energy organization in the q subspace. Figure 4B projects the coordinate-like latent q into two principal components and colors each state by $H_\varphi(q,p)$. The projected trajectory exhibits a clear energy-ordered structure: high-energy states are concentrated near one end of the curve, while lower-energy states occupy a different region of the projected subspace. This smooth variation indicates that the learned Hamiltonian is systematically coupled to the q component of the phase representation. Therefore, the energy term used by HTC^2 is computed over an organized phase subspace rather than over an unstructured latent cloud.

(C) Energy organization in the p subspace. Figure 4C shows the corresponding projection for the momentum-like latent p. The learned energy again varies smoothly across the projected trajectory, with a compact low-energy region and a separate high-energy branch. This indicates that $H_\varphi(q,p)$ depends on both components of the learned phase state. Together, the q and p visualizations support the design of HTC^2: temporal future discovery is performed in a Hamiltonian phase space whose energy varies coherently over the learned latent dynamics, and whose controlled evolution is modulated by the policy actions.

We further analyze the relation between the q and p subspaces in Appendix C.3. The dimension-wise coupling maps show that q and p are dynamically related but not uniformly coupled: only selected cross-dimensional pairs exhibit strong correlations. This supports a non-collapsed phase representation in which coordinate-like and momentum-like variables share the same trajectory dynamics while preserving distinct representational roles.

4.5 Parametric OOD Robustness

We further evaluate the zero-shot robustness of SAC policies trained with HTC^2 under parametric dynamics shifts. The final policies trained on the original DMC environments are evaluated on variants of `walker_walk`, `hopper_stand`, and `cheetah_run` with modified gravity ($1.5\times$), friction ($2.0\times$), mass ($1.5\times$), or damping ($2.0\times$). We report return retention, defined as the OOD return divided by the corresponding in-distribution return for the same method and seed.

Figure 5 shows that HTC^2 maintains high retention under several physical shifts. On `walker_walk`, HTC^2 achieves the highest retention under gravity, friction, and mass perturbations, with an average retention of 89.7% across the four variants. On `hopper_stand`, it performs strongly under the mass shift and remains competitive under friction and damping, while RIDE obtains the highest retention under the gravity shift. On `cheetah_run`, HTC^2 achieves the highest retention under gravity and mass perturbations, with mass-shift retention exceeding 100%.

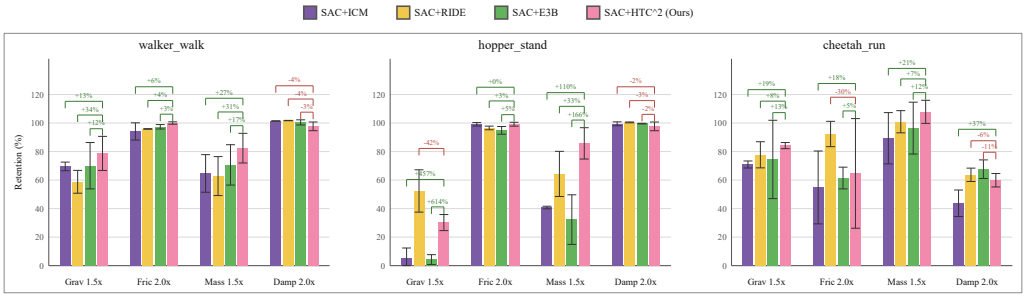


Figure 5: **Zero-shot policy retention under parametric physical shifts.** Bars and error bars show mean $\pm$ standard deviation over four random seeds. Higher values indicate greater return retention.

Across all three tasks and four parameter shifts, HTC^2 obtains the highest overall average retention (82.3%), compared with RIDE (80.6%), E3B (72.5%), and ICM (69.4%). Detailed per-shift results are reported in Appendix C.4.

4.6 Ablation Study

We ablate the main components of HTC^2 on `walker_walk`, `hopper_hop`, and `cheetah_run`. Specifically, w/o $\mathcal{L}_{\text{phys}}$ removes the physical-consistency loss; w/o G_H removes the Hamiltonian modulation; one-step ΔH replaces the temporal contrastive reward with one-step Hamiltonian change; w/o RND removes the warm-start schedule; and w/o repr. norm disables contrastive representation normalization. All variants use SAC with the same training budget and are evaluated over four random seeds.

Table 1: **HTC^2 ablations.** Final return (mean $\pm$ std, four seeds).

Variant	Walker	Hopper	Cheetah
w/o $\mathcal{L}_{\text{phys}}$	693.166 ± 203.967	245.416 ± 14.651	566.209 ± 159.871
w/o G_H	718.941 ± 16.122	235.203 ± 1.896	609.108 ± 16.185
one-step ΔH	761.051 ± 195.324	1.843 ± 1.183	510.093 ± 128.266
w/o RND	785.069 ± 93.809	227.774 ± 0.719	578.816 ± 9.910
w/o repr. norm	757.758 ± 114.117	144.466 ± 85.169	594.697 ± 16.876
HTC^2 full	$\mathbf{814.342 \pm 162.857}$	$\mathbf{250.865 \pm 9.598}$	$\mathbf{659.810 \pm 33.525}$

Table 1 shows that the full HTC^2 configuration achieves the highest mean final return on all three tasks. Removing $\mathcal{L}_{\text{phys}}$ or G_H lowers the mean return across all environments, with larger reductions from removing $\mathcal{L}_{\text{phys}}$ on `walker_walk` and `cheetah_run`. Replacing the temporal contrastive reward with one-step ΔH produces the largest reduction on `hopper_hop`, where the final return drops to 1.843 ± 1.183.

Removing the RND warm start or representation normalization also reduces performance across all three tasks. The reduction from removing representation normalization is particularly pronounced on `hopper_hop`, whereas removing the RND warm start produces a larger absolute decrease on `cheetah_run`. Across the three tasks, the complete configuration yields the strongest mean final performance. Full learning curves are reported in Appendix C.5.

5 Conclusion

We presented **Hamiltonian Temporal Contrastive Curiosity** (**HTC^2**), an intrinsic exploration method for discovering temporally informative futures in a learned Hamiltonian phase space. HTC^2 uses a horizon-conditioned temporal contrastive critic to identify poorly organized multi-step futures and modulates this signal with bounded Hamiltonian variation to emphasize structured changes in the learned phase dynamics. Across DeepMind Control Suite tasks, HTC^2 shows strong control performance with both SAC and PPO, competitive physical-state coverage, and robust retention under parametric dynamics shifts. Ablations show performance degradation when removing physical consistency, Hamiltonian modulation, warm start, or representation normalization, while the single-step Hamiltonian-change reward performs substantially worse on `hopper_hop`. Hamiltonian diagnostics further reveal energy-organized phase representations with distinct responses to policy-induced control. Overall, these results characterize HTC^2 as a physics-structured approach to long-horizon exploration in physical control.

AI USE STATEMENT

During the preparation of this manuscript, the authors used ChatGPT to assist with language polishing, including grammar correction and phrasing improvements. ChatGPT was not used for research ideation. The authors reviewed all AI-assisted text and take full responsibility for the final content of this work.

REFERENCES

Amirhossein Arzani, Lingxiao Yuan, Pania Newell, and Bei Wang. Interpreting and generalizing deep learning in physics-based problems with functional linear models. *Engineering with Computers*, 41(1):135–157, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

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

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018a.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018b.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

Congyue Deng, Brandon Y Feng, Cecilia Garraffo, Alan Garbarz, Robin Walters, William T Freeman, Leonidas Guibas, and Kaiming He. Denoising hamiltonian network for physical reasoning. *arXiv preprint arXiv:2503.07596*, 2025.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. *arXiv preprint arXiv:2011.08909*, 2020.

Alejandro Castaneda Garcia, Jan Warchocki, Jan Van Gemert, Daan Brinks, and Nergis Tomen. Learning physics from video: Unsupervised physical parameter estimation for continuous dynamical systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27924–27933, 2025.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

Zhaohan Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35:31855–31870, 2022.

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

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, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, pp. 4116–4126. PMLR, 2020.

Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via elliptical episodic bonuses. *Advances in Neural Information Processing Systems*, 35:37631–37646, 2022.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Yuren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. *Advances in Neural Information Processing Systems*, 36: 31831–31864, 2023.

Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu, Yizhou Wang, and He Wang. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11058–11068. IEEE, 2025.

Augustine Mavor-Parker, Kimberly Young, Caswell Barry, and Lewis Griffin. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In *International conference on machine learning*, pp. 15220–15240. PMLR, 2022.

Faisal Mohamed, Catherine Ji, Benjamin Eysenbach, and Glen Berseth. Temporal representations for exploration: Learning complex exploratory behavior without extrinsic rewards. *arXiv preprint arXiv:2603.02008*, 2026.

Kohei Noda, Araki Wakiuchi, Yoshihiro Hayashi, and Ryo Yoshida. Advancing extrapolative predictions of material properties through learning to learn using extrapolative episodic training. *Communications Materials*, 6(1):36, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pp. 2721–2730. PMLR, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pp. 5062–5071. PMLR, 2019.

Roberta Raileanu and Tim Rocktäschel. Ride: Rewarding impact-driven exploration for procedurally-generated environments. *arXiv preprint arXiv:2002.12292*, 2020.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pp. 5628–5637. PMLR, 2019.

Robin Schiewer, Anand Subramoney, and Laurenz Wiskott. Exploring the limits of hierarchical world models in reinforcement learning. *Scientific Reports*, 14(1):26856, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pp. 8583–8592. PMLR, 2020.

Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Jinzhou Tang, Fan Feng, Minghao Fu, Wenjun Lin, Jing Yang, Biwei Huang, and Keze Wang. Dreamsac: Learning hamiltonian world models via symmetry exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15188–15198, 2026.

Sebastian B Thrun. *Efficient exploration in reinforcement learning*. Carnegie Mellon University, 1992.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. *arXiv preprint arXiv:1909.13789*, 2019.

Arne Troch, Kevin Mets, and Siegfried Mercelis. Action-conditioned hamiltonian generative networks (ac-hgn) for supervised and reinforcement learning. In *7th Annual Learning for Dynamics & Control Conference, 04-06 June, 2025, Ann Arbor, Michigan, USA*, pp. 310–322, 2025.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yingchen Xu, Jack Parker-Holder, Aldo Pacchiano, Philip Ball, Oleh Rybkin, Stephen Roberts, Tim Rocktäschel, and Edward Grefenstette. Learning general world models in a handful of reward-free deployments. *Advances in Neural Information Processing Systems*, 35:26820–26838, 2022.

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.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. *arXiv preprint arXiv:1909.12077*, 2019.

Alicja Ziarko, Michał Bortkiewicz, Michał Zawalski, Benjamin Eysenbach, and Piotr Miłoś. Contrastive representations for temporal reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=UlkcH5Ccrk.

A Detailed Methodology

This section provides additional derivations for the controlled Hamiltonian relations introduced in Equation 1 and 2, followed by the leapfrog discretization used by the Hamiltonian transition core. We then derive a continuous-time interpretation of the multi-step Hamiltonian variation defined in Equation 12.

A.1 Controlled Hamiltonian Dynamics and Energy Balance

Autonomous Hamiltonian conservation. Starting from the controlled dynamics in Equation 1, consider first the autonomous case $a_t = 0$. For a Hamiltonian with no explicit time dependence, its derivative along the trajectory is

$$\begin{aligned}\frac{dH_\varphi}{dt} &= \left(\frac{\partial H_\varphi}{\partial q}\right)^\top \dot{q} + \left(\frac{\partial H_\varphi}{\partial p}\right)^\top \dot{p} \\ &= \left(\frac{\partial H_\varphi}{\partial q}\right)^\top \frac{\partial H_\varphi}{\partial p} - \left(\frac{\partial H_\varphi}{\partial p}\right)^\top \frac{\partial H_\varphi}{\partial q} \\ &= 0.\end{aligned} \tag{19}$$

This is the standard conservation relation of an autonomous Hamiltonian system. In classical mechanics, it is associated with time-translation invariance through Noether's theorem: in the absence of explicit time dependence and external forcing, the Hamiltonian remains constant along the continuous trajectory.

Energy balance under control. Restoring the action-dependent term in Equation 1 gives

$$\begin{aligned}\frac{dH_\varphi}{dt} &= \left(\frac{\partial H_\varphi}{\partial q}\right)^\top \frac{\partial H_\varphi}{\partial p} \\ &\quad + \left(\frac{\partial H_\varphi}{\partial p}\right)^\top \left(-\frac{\partial H_\varphi}{\partial q} + g_\varphi(q)a_t\right) \\ &= \left(\frac{\partial H_\varphi}{\partial p}\right)^\top g_\varphi(q)a_t.\end{aligned} \tag{20}$$

Using $\dot{q} = \partial H_\varphi/\partial p$, the last term can equivalently be written as $\dot{q}^\top g_\varphi(q)a_t$. Thus, Equation 20 recovers the controlled energy-balance relation in Equation 2.

A.2 Symplectic Integration of the Controlled Dynamics

The controlled continuous-time dynamics in Equation 1 must be discretized to construct the Hamiltonian transition prior. Our implementation uses a leapfrog integrator with step size $\Delta t = 0.1$, with a_t held fixed over one integration step.

The update follows the kick–drift–kick structure. Starting from (q_t, p_t), the first momentum half-step is

$$p_{t+\frac{1}{2}} = p_t - \frac{\Delta t}{2}\frac{\partial H_\varphi}{\partial q}(q_t, p_t) + \frac{\Delta t}{2} g_\varphi(q_t)a_t. \tag{21}$$

The intermediate momentum-like state is then used to advance the coordinate-like variable over a full step:

$$q_{t+1} = q_t + \Delta t \frac{\partial H_\varphi}{\partial p}\left(q_t, p_{t+\frac{1}{2}}\right). \tag{22}$$

Finally, the momentum update is completed at the new position:

$$p_{t+1} = p_{t+\frac{1}{2}} - \frac{\Delta t}{2}\frac{\partial H_\varphi}{\partial q}\left(q_{t+1}, p_{t+\frac{1}{2}}\right) + \frac{\Delta t}{2} g_\varphi(q_{t+1})a_t. \tag{23}$$

The three updates constitute the Hamiltonian physics step used in the world model. Gradients of H_φ with respect to q and p are evaluated through automatic differentiation, while the learned control term is incorporated into both momentum half-steps.

For the autonomous Hamiltonian component, leapfrog provides a symplectic discretization and is designed to preserve the geometric structure of Hamiltonian dynamics under repeated integration. Symplectic integration does not imply exact conservation of H_φ at every discrete step. The complete world-model prior also contains the learned control contribution and a stochastic residual parameterization; accordingly, the leapfrog step is used as a structure-preserving inductive bias for the Hamiltonian transition core.

A.3 Temporal Hamiltonian Variation

The controlled energy-balance relation in Equation 2 provides a continuous-time interpretation of the multi-step Hamiltonian variation used by HTC^2. Integrating it over the interval $[t, t+\tau]$ gives

$$H_\varphi(z_{t+\tau}) - H_\varphi(z_t) = \int_t^{t+\tau} \left(\frac{\partial H_\varphi}{\partial p}(z_s)\right)^\top g_\varphi(q_s) a_s \, ds. \tag{24}$$

For a sampled discrete horizon k, HTC^2 uses the endpoint magnitude $\Delta H_{t,k}$ defined in Equation 12. Hence, the multi-step term summarizes the Hamiltonian variation across the reached temporal interval rather than only an adjacent one-step change.

HTC^2 does not optimize this quantity independently. The endpoint variation is transformed by the bounded gate in Equation 13 and modulates the temporal mismatch $D_\theta(t,k)$ defined in Equation 11. The resulting intrinsic signal is given in Equation 14. Thus, temporal future discovery remains the primary exploration signal, while Hamiltonian variation provides a bounded structural modulation over the same reached multi-step transition.

B Implementation Details

We provide implementation details of the Hamiltonian latent world model, its optimization objectives, the temporal contrastive critic, and the intrinsic-reward schedule used in HTC^2. Algorithms 1 and 2 summarize the integration of HTC^2 with SAC and PPO, respectively.

B.1 Hamiltonian Latent World Model

At time t, the recurrent latent state consists of a deterministic state h_t and a phase representation

$$z_t = [q_t, p_t], \tag{25}$$

where q_t and p_t denote coordinate-like and momentum-like latent variables. The deterministic state is updated by the recurrent core,

$$h_{t+1} = \mathrm{GRU}\left(F_{\mathrm{rnn}}([q_t, p_t, a_t]), h_t\right), \tag{26}$$

where F_{rnn} denotes the input projection preceding the GRU.

Let

$$c_t = E_\varphi(x_t) \tag{27}$$

denote the observation token produced by the visual encoder E_φ. The posterior network predicts categorical distributions over the two phase components:

$$Q_{q,t} = q_\varphi(q_t \mid h_t, c_t), \tag{28}$$
$$Q_{p,t} = q_\varphi(p_t \mid h_t, c_t). \tag{29}$$

Their continuous posterior expectations are

$$\mu_{q,t} = \mathbb{E}_{Q_{q,t}}[q_t], \qquad \mu_{p,t} = \mathbb{E}_{Q_{p,t}}[p_t]. \tag{30}$$

For compactness, q_t and p_t below denote these continuous phase values whenever the Hamiltonian transition or temporal critic is evaluated.

The Hamiltonian transition follows the controlled dynamics in Equation 1. The learned input mapping

$$g_\varphi(q_t) \in \mathbb{R}^{d_p \times d_a} \tag{31}$$

maps the action $a_t \in \mathbb{R}^{d_a}$ to the momentum-like control term $g_\varphi(q_t)a_t \in \mathbb{R}^{d_p}$. Using the leapfrog updates in Equation 21 and 23, one deterministic Hamiltonian step is written compactly as

$$(\bar{q}_{t+1}, \bar{p}_{t+1}) = \Phi_{\Delta t}^{\mathrm{LF}}\left(q_t, p_t, a_t; H_\varphi, g_\varphi\right), \tag{32}$$

where $\Phi_{\Delta t}^{\mathrm{LF}}$ denotes one controlled leapfrog step and $\Delta t = 0.1$.

The continuous prediction is subsequently mapped to categorical prior logits. We denote by $\mathcal{G}_{\sigma_\ell}(\cdot)$ the deterministic Gaussian-shaped logit projection with width parameter σ_ℓ. The prior logits are

$$\ell_{q,t+1}^{\mathrm{prior}} = \mathcal{G}_{\sigma_\ell}(\bar{q}_{t+1}) + R_q(\chi_t), \tag{33}$$

$$\ell_{p,t+1}^{\mathrm{prior}} = \mathcal{G}_{\sigma_\ell}(\bar{p}_{t+1}) + R_p(\chi_t), \tag{34}$$

where

$$\chi_t = [h_{t+1}, q_t, p_t, a_t]. \tag{35}$$

For $\nu \in \{q, p\}$, the gated residual branch is

$$R_\nu(\chi) = s\,\sigma(G_\nu(\chi)) \odot \tanh(F_\nu(\chi))\,, \tag{36}$$

where F_ν and G_ν denote the residual-value and gate networks, respectively, $\sigma(\cdot)$ is the sigmoid function, $\odot$ denotes element-wise multiplication, and s is the residual scale. We use $s = 1.0$. The residual output is zero-initialized and the gate is initialized negatively so that the Hamiltonian prediction dominates early training.

The categorical prior distributions used below are obtained as

$$P_{q,t+1} = \mathrm{Cat}\left(\mathrm{softmax}(\ell_{q,t+1}^{\mathrm{prior}})\right), \tag{37}$$

$$P_{p,t+1} = \mathrm{Cat}\left(\mathrm{softmax}(\ell_{p,t+1}^{\mathrm{prior}})\right). \tag{38}$$

B.2 World Model Loss Functions

The complete world-model objective is defined in Equation 7. We use $\alpha = 0.8$ and $\lambda_{\mathrm{phys}} = 1.0$. This section specifies the individual loss terms used to optimize the posterior representation and the Hamiltonian dynamics prior.

Reconstruction Loss ($\mathcal{L}_{\mathrm{rec}}$). Define the posterior state

$$s_t^{\mathrm{post}} = [h_t, \mu_{q,t}, \mu_{p,t}]. \tag{39}$$

The decoder D_φ reconstructs the next observation from the posterior state:

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}\left[\left\|D_\varphi(s_{t+1}^{\mathrm{post}}) - x_{t+1}\right\|_2^2\right]. \tag{40}$$

Dynamics and Representation Losses ($\mathcal{L}_{\mathrm{dyn}}, \mathcal{L}_{\mathrm{rep}}$). The dynamics loss updates the prior while stopping gradients through the posterior distributions:

$$\begin{aligned}\mathcal{L}_{\mathrm{dyn}} = \mathbb{E}\Big[&D_{\mathrm{KL}}\left(\mathrm{sg}[Q_{q,t+1}] \| P_{q,t+1}\right) \\ &+ D_{\mathrm{KL}}\left(\mathrm{sg}[Q_{p,t+1}] \| P_{p,t+1}\right)\Big].\end{aligned} \tag{41}$$

Conversely, the representation loss updates the posterior while stopping gradients through the prior:

$$\begin{aligned}\mathcal{L}_{\mathrm{rep}} = \mathbb{E}\Big[&D_{\mathrm{KL}}\left(Q_{q,t+1} \| \mathrm{sg}[P_{q,t+1}]\right) \\ &+ D_{\mathrm{KL}}\left(Q_{p,t+1} \| \mathrm{sg}[P_{p,t+1}]\right)\Big].\end{aligned} \tag{42}$$

Here, $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. In the implementation, the q- and p-KL terms are summed before applying standard free-nats stabilization and batch averaging.

Inverse-Dynamics Loss ($\mathcal{L}_{\mathrm{inv}}$). The inverse-dynamics model predicts the executed action from adjacent posterior states,

$$\hat{a}_t = F_{\mathrm{inv}}\left([s_t^{\mathrm{post}}, s_{t+1}^{\mathrm{post}}]\right), \tag{43}$$

and is optimized by

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}\left[\|\hat{a}_t - a_t\|_2^2\right]. \tag{44}$$

Hamiltonian Consistency Loss ($\mathcal{L}_{\text{phys}}$). The continuous Hamiltonian prediction from Equation 32 is aligned with the posterior phase expectation:

$$\mathcal{L}_{\text{phys}} = \mathbb{E}\Big[\|\bar{q}_{t+1} - \text{sg}[\mu_{q,t+1}]\|_2^2 + \|\bar{p}_{t+1} - \text{sg}[\mu_{p,t+1}]\|_2^2\Big]. \tag{45}$$

B.3 Temporal Contrastive Critic

The temporal critic operates on posterior phase trajectories inferred from collected experience. Future horizons are sampled according to the truncated geometric distribution in Equation 4, with $K = 16$ and $\gamma_{\text{T}} = 0.98$ in the reported configuration. Only anchor–future pairs belonging to the same episode are retained.

The anchor and future representations follow Equation 8. Both representations are ℓ_2-normalized before evaluating the score in Equation 9. The critic uses a representation dimension of 64 and a horizon-embedding dimension of 16. The temperature is initialized to 0.1 and optimized jointly with the critic.

The temporal critic is trained using $\mathcal{L}_{\text{TC}}$ in Equation 10 with a batch size of 256 and AdamW learning rate 3×10^{-4}. Gradients are stopped at the phase-state input in the default configuration, so this objective updates the temporal critic without directly propagating gradients into the Hamiltonian world model.

At reward computation time, the temporal mismatch is evaluated using Equation 11, while the Hamiltonian modulation and the raw HTC^2 reward are computed according to Equation 13 and 14, respectively. The HTC^2 reward is then normalized using expected-distance normalization. For ℓ_2-normalized representations and Euclidean distance, the reference distance is $\sqrt{2}$.

B.4 Cold Start and Reward Scheduling

The cold-start RND (Burda et al., 2018b) module operates on the posterior phase representation z_t. It consists of a trainable predictor f_ξ and a fixed randomly initialized target network $\bar{f}$. The raw RND signal is

$$r_{\text{RND}}^{\text{raw}}(t) = \left\| f_\xi(z_t) - \bar{f}(z_t) \right\|_2^2. \tag{46}$$

The predictor and target are MLPs with hidden dimension 256 and output dimension 128. The predictor is optimized with AdamW at learning rate 3×10^{-4}.

The RND signal is normalized by running statistics before reward mixing. The RND-to-HTC^2 transition follows Equation 15] and 16, with $T_{\text{warm}} = 50{,}000$ and $T_{\text{anneal}} = 200{,}000$ environment steps. The resulting intrinsic reward is combined with the extrinsic reward according to Equation 17.

B.5 Hyperparameters

The core exploration hyperparameters are summarized in Table 2, while the policy-optimizer configurations are reported in Table 3.

B.6 Training Pseudocode

Algorithms 1 and 2 summarize the integration of HTC^2 with off-policy SAC and on-policy PPO, respectively. Standard SAC and PPO optimization steps are omitted for clarity.

C Additional Experiments and Diagnostics

C.1 Detailed DMC Control Results

This section provides the complete numerical results and learning curves corresponding to the DMC control benchmark in Section 4.2. All reported final returns are computed over five random seeds and are shown as mean $\pm$ standard deviation. Table 4 reports the final control performance for ICM, RIDE, E3B, and HTC^2 under SAC and PPO, with Plan2Explore additionally included in the PPO comparison. Figures 6 and 7 provide the corresponding SAC and PPO learning curves.

Table 2: Core hyperparameters for HTC^2 and intrinsic-reward baselines.

HTC^2		**Intrinsic-reward baselines**			
Hyperparameter	Value	Hyperparameter	ICM	E3B	RIDE
Intrinsic scale β	0.05	Intrinsic scale β	0.05	0.05	0.05
Cold-start reward	RND	Latent dim	256	256	256
Warmup steps	50K	Reward lr	10^{-3}	10^{-3}	10^{-3}
Annealing steps	200K	Batch size	256	256	256
World-model lr	3×10^{-4}	Update proportion	1.0	1.0	1.0
World-model batch size	256	Reward normalization	RMS	RMS	RMS
Deterministic state dim	256	E3B ridge	–	0.1	–
Stochastic latent dim	32	RIDE k	–	–	10
q / p dim	16 / 16	Kernel cluster distance	–	–	0.008
Categorical classes	32	Kernel epsilon	–	–	10^{-4}
Sequence chunk	32	RIDE c	–	–	0.001
Integrator / Δt	Leapfrog / 0.1	RIDE s_m	–	–	8.0
Logit projection std.	0.1	κ	0.0	0.0	0.0
Residual scale	1.0	γ	–	–	–
KL balance α	0.8				
Free-nats stabilization	1.0				
Physics loss weight	1.0				
Temporal horizon k	1–16				
Horizon sampling	Geometric (0.98)				
Temporal critic lr / batch	3×10^{-4} / 256				
Critic hidden / repr. dim	256 / 64				
Horizon embedding dim	16				
Temperature init.	0.1 (learned)				
Similarity	Negative ℓ_2				
Representation normalization	Yes				
Energy gate α_H / c_H	0.1 / 5.0				
Smoothness λ_{sm}	0.01				

Table 3: SAC and PPO hyperparameters.

Hyperparameter	SAC	PPO
Learning rate	1×10^{-4}	1.5×10^{-5}
Feature dim	50	512
Hidden dim	1024	1024
Batch size	256	256
Discount γ	0.99	0.99
Initial random steps	5000	–
Replay size	50000	–
Initial temperature	0.1	–
Target update τ	0.005	–
Actor / target update freq.	1 / 2	–
Rollout steps	–	256
PPO epochs	–	15
Clip range / value clip	–	0.2 / 0.1
Entropy coefficient	–	0.1
Value coefficient	–	0.5
Max gradient norm	–	0.5

With SAC, HTC^2 achieves the highest mean final return on five of the six tasks. On `walker_stand`, HTC^2 obtains 966.5 ± 3.8, compared with 967.1 ± 3.3 for RIDE. The largest absolute difference relative to the strongest competing baseline occurs on `walker_walk`, where HTC^2 reaches 827.0 ± 115.2, compared with 717.9 ± 67.5 for RIDE.

Algorithm 1 HTC^2 with Soft Actor–Critic

Params: $\pi_\eta, Q_{\omega_1}, Q_{\omega_2}, p_\varphi, (\phi_\theta, \psi_\theta), f_\xi$ ▷ policy, critics, world model, TC critic, RND
Init: replay buffer $\mathcal{D}$, target critics, fixed RND target $\bar{f}$ ▷ initialization
while not converged **do**
  $a_t \sim \pi_\eta(\cdot \mid x_t), \quad \mathcal{D} \leftarrow \mathcal{D} \cup \{x_t, a_t, r_t^{\text{ext}}, x_{t+1}, d_t\}$ ▷ environment interaction
  $\mathcal{B}_{\text{seq}} \sim \mathcal{D}, \quad z_{1:T} \leftarrow q_\varphi^{\text{post}}(\mathcal{B}_{\text{seq}})$ ▷ posterior phase trajectory
  $k_i \sim \omega_k, \quad z_i^+ \leftarrow z_{t_i+k_i}$ ▷ Equation 4
  $\theta \leftarrow \theta - \eta_{\text{TC}} \nabla_\theta \mathcal{L}_{\text{TC}}$ ▷ Equation 10
  $r_i^{\text{HTC}^2} \leftarrow \text{Norm}\left(D_\theta(t_i, k_i) G_H(t_i, k_i) - \lambda_{\text{sm}} \|a_{t_i} - a_{t_i-1}\|_2^2\right)$ ▷ Equation 14
  $r_i^{\text{RND}} \leftarrow \text{Norm}\left(\|f_\xi(z_{t_i}) - \bar{f}(z_{t_i})\|_2^2\right)$ ▷ latent RND
  $w_n \leftarrow \text{clip}\left(\frac{n - T_{\text{warm}}}{T_{\text{anneal}}}, 0, 1\right)$ ▷ Equation 15
  $r_i^{\text{int}} \leftarrow (1 - w_n) r_i^{\text{RND}} + w_n r_i^{\text{HTC}^2}$ ▷ Equation 16
  $\tilde{r}_i \leftarrow r_i^{\text{ext}} + \beta r_i^{\text{int}}$ ▷ Equation 17
  $\mathcal{B}_{\text{RL}} \leftarrow \{(x_{t_i}, a_{t_i}, \tilde{r}_i, x_{t_i+1}, d_{t_i})\}_{i=1}^B$ ▷ anchor transitions
  $(\eta, \omega_1, \omega_2) \leftarrow \text{SACUpdate}(\mathcal{B}_{\text{RL}})$ ▷ standard SAC update
  $\varphi \leftarrow \varphi - \eta_{\text{WM}} \nabla_\varphi \mathcal{L}_{\text{WM}}$ ▷ Equation 7
  $\xi \leftarrow \xi - \eta_{\text{RND}} \nabla_\xi \|f_\xi(z) - \bar{f}(z)\|_2^2$ ▷ RND predictor update

Algorithm 2 HTC^2 with Proximal Policy Optimization

Params: $\pi_\eta, V_\omega, p_\varphi, (\phi_\theta, \psi_\theta), f_\xi$ ▷ policy, value, world model, TC critic, RND
Init: rollout buffer $\mathcal{R}$, fixed RND target $\bar{f}$ ▷ initialization
for each rollout iteration **do**
  $\mathcal{R} \leftarrow \{x_t, a_t, r_t^{\text{ext}}, x_{t+1}, d_t\}_{t=1}^{T_{\text{roll}}}$ ▷ on-policy rollout
  $z_{1:T} \leftarrow q_\varphi^{\text{post}}(\mathcal{R})$ ▷ posterior phase trajectory
  $k_i \sim \omega_k, \quad z_i^+ \leftarrow z_{t_i+k_i}$ ▷ Equation 4
  $\theta \leftarrow \theta - \eta_{\text{TC}} \nabla_\theta \mathcal{L}_{\text{TC}}$ ▷ Equation 10
  $r_i^{\text{HTC}^2} \leftarrow \text{Norm}\left(D_\theta(t_i, k_i) G_H(t_i, k_i) - \lambda_{\text{sm}} \|a_{t_i} - a_{t_i-1}\|_2^2\right)$ ▷ Equation 14
  $r_i^{\text{RND}} \leftarrow \text{Norm}\left(\|f_\xi(z_{t_i}) - \bar{f}(z_{t_i})\|_2^2\right)$ ▷ latent RND
  $w_n \leftarrow \text{clip}\left(\frac{n - T_{\text{warm}}}{T_{\text{anneal}}}, 0, 1\right)$ ▷ Equation 15
  $r_i^{\text{int}} \leftarrow (1 - w_n) r_i^{\text{RND}} + w_n r_i^{\text{HTC}^2}$ ▷ Equation 16
  $\tilde{r}_i \leftarrow r_i^{\text{ext}} + \beta r_i^{\text{int}}$ ▷ Equation 17
  $(\hat{A}_i, R_i) \leftarrow \text{GAE}(\mathcal{R}, \tilde{r}, V_\omega)$ ▷ advantages and returns
  $(\eta, \omega) \leftarrow \text{PPOUpdate}(\mathcal{R}, \hat{A}, R)$ ▷ standard clipped PPO update
  $\varphi \leftarrow \varphi - \eta_{\text{WM}} \nabla_\varphi \mathcal{L}_{\text{WM}}$ ▷ Equation 7
  $\xi \leftarrow \xi - \eta_{\text{RND}} \nabla_\xi \|f_\xi(z) - \bar{f}(z)\|_2^2$ ▷ RND predictor update

Under PPO, HTC^2 achieves the highest mean final return on all six tasks among the evaluated methods, including Plan2Explore. On `walker_stand`, HTC^2 reaches 930.6 ± 20.1, while Plan2Explore obtains 929.2 ± 6.2, showing very similar mean performance in this high-return regime. On `hopper_hop`, HTC^2 obtains 100.7 ± 58.8, compared with 89.5 ± 61.2 for the strongest competing baseline, ICM.

Overall, the SAC and PPO results show that the strong mean final-return performance of HTC^2 spans multiple control tasks rather than being driven by a single environment. The complete learning curves further show how these differences evolve over training.

C.2 Detailed Physical-State Coverage Results

We provide the detailed physical-state discretization and numerical statistics corresponding to Figure 3. All statistics are computed over four random seeds and reported as mean $\pm$ standard deviation.

Coverage-state construction. For `cartpole_swingup`, we use

$$s_t^{\text{phy}} = [x_{\text{cart},t}, \text{wrap}(\theta_{\text{pole},t})], \tag{47}$$

Table 4: **Final control performance on six DMC tasks.** Episode return over five random seeds, reported as mean ± standard deviation. Within each optimizer block and task, red, orange, and yellow denote the first, second, and third-ranked exploration methods, respectively. Higher is better.

RL	Exploration	Walker Walk	Walker Run	Walker Stand	Hopper Hop	Hopper Stand	Cheetah Run
SAC	ICM	677.8±230.2	356.7±52.9	936.9±31.7	237.5±9.1	863.3±29.5	631.9±45.4
	RIDE	717.9±67.5	364.9±117.7	967.1±3.3	239.0±11.3	871.3±22.1	624.7±44.0
	E3B	679.2±199.4	401.6±92.1	961.3±8.1	235.4±24.7	858.0±22.1	619.5±40.2
	HTC^2 (ours)	827.0±115.2	415.8±111.4	966.5±3.8	248.8±6.1	883.8±12.5	656.5±45.0
PPO	ICM	601.9±179.6	269.1±47.1	921.0±25.4	89.5±61.2	637.6±55.6	451.5±69.2
	RIDE	641.1±153.5	254.2±23.1	877.4±82.0	83.4±68.5	565.8±174.0	425.8±43.8
	E3B	698.5±129.7	277.5±16.4	865.7±100.1	55.3±65.4	628.1±53.9	445.9±33.3
	Plan2Explore	647.8±47.3	239.2±38.4	929.2±6.2	74.2±68.6	620.7±51.8	435.2±37.6
	HTC^2 (ours)	733.6±160.7	307.9±22.8	930.6±20.1	100.7±58.8	670.2±25.1	458.0±27.4

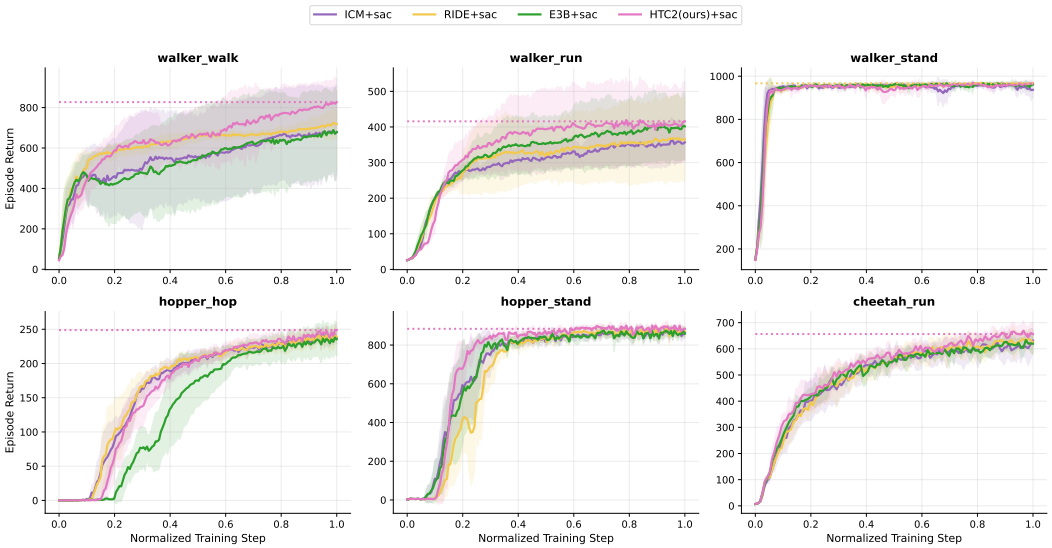


Figure 6: **SAC learning curves on six DMC tasks.** Curves show the mean episode return over five random seeds, with shaded regions indicating one standard deviation, for SAC combined with ICM, RIDE, E3B, and HTC^2.

with a 60×72 grid over $x_{\text{cart}} \in [-0.50, 0.50]$ and $\theta_{\text{pole}} \in [-\pi, \pi]$.

For `ball_in_cup_catch`, the physical state is

$$s_t^{\text{phy}} = [x_{\text{ball},t} - x_{\text{cup},t}, z_{\text{ball},t} - z_{\text{cup},t}], \tag{48}$$

which is discretized using a 64×64 grid, with both dimensions restricted to $[-0.35, 0.35]$.

For `finger_spin`, we use

$$s_t^{\text{phy}} = \left[\text{wrap}(\theta_{\text{obj},t}), \text{clip}(\dot{\theta}_{\text{obj},t}, -25, 25)\right], \tag{49}$$

with a 72×80 grid over $\theta_{\text{obj}} \in [-\pi, \pi]$ and $\dot{\theta}_{\text{obj}} \in [-25, 25]$. These task-specific physical states are used only for evaluation and do not affect policy learning.

Table 5 reports the complete coverage and reward statistics for the three environments. On `cartpole_swingup`, HTC^2 reaches 4320.0 ± 0.0 visited states at the end of training and obtains a coverage AUC of 3525.7 ± 71.6, both the highest mean values among the four methods. Its final episode reward and reward AUC are 579.6 ± 273.1 and 134.7 ± 87.9, respectively, which are also the highest reported means for this task.

On `ball_in_cup_catch`, RIDE achieves the highest final coverage, 2731.5 ± 14.8, followed by ICM at 2680.5 ± 3.5, E3B at 2664.5 ± 12.0, and HTC^2 at 2655.0 ± 55.2. In terms of coverage AUC, HTC^2 obtains the second-highest mean value of 2415.1 ± 30.8, behind RIDE at 2499.9 ± 14.4. For downstream control, HTC^2 reaches a final reward of 958.7 ± 2.4 and a reward AUC of 708.6 ± 87.1,

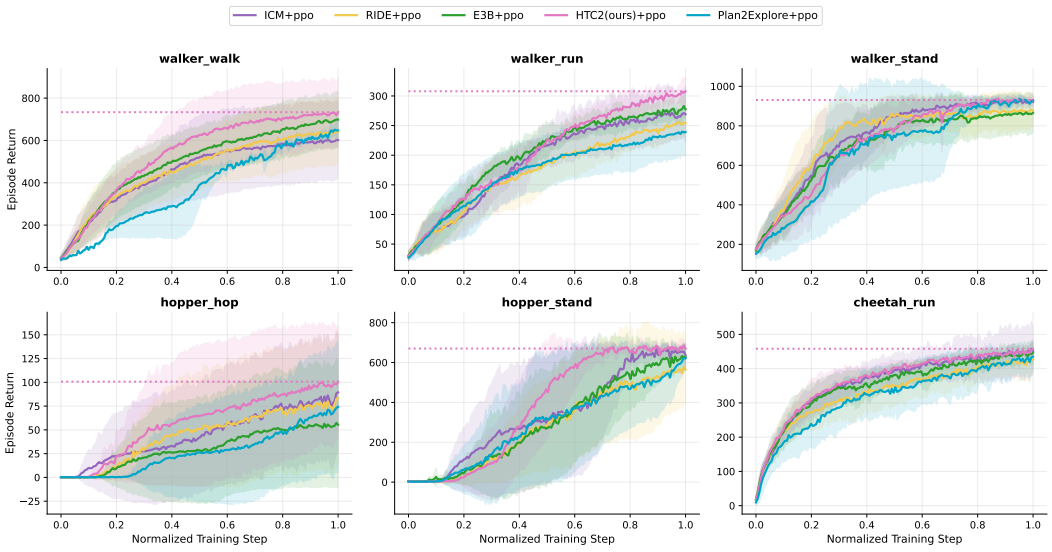


Figure 7: **PPO learning curves on six DMC tasks.** Curves show the mean episode return over five random seeds, with shaded regions indicating one standard deviation, for PPO combined with ICM, RIDE, E3B, Plan2Explore, and HTC^2.

Table 5: **Physical-state coverage and downstream control performance on three DMC environments.** Final performance and area under the training curve (AUC) are reported as mean $\pm$ standard deviation over four random seeds. For each row, the best and second-best mean values are highlighted in green and yellow, respectively. The highlighting reflects ranking by the reported means. Higher values are better for all metrics.

Environment	Metric	Statistic	HTC^2 (Ours) + SAC	ICM + SAC	RIDE + SAC	E3B + SAC
Cartpole Swingup	Visited States	Final	4320.000 $\pm$ 0.000	3862.250 $\pm$ 486.344	3759.500 $\pm$ 249.116	3384.750 $\pm$ 399.951
		AUC	3525.652 $\pm$ 71.624	3187.527 $\pm$ 163.460	3200.982 $\pm$ 70.494	2976.313 $\pm$ 208.753
	Episode Reward	Final	579.550 $\pm$ 273.080	126.855 $\pm$ 253.710	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000
		AUC	134.662 $\pm$ 87.884	15.452 $\pm$ 30.791	0.058 $\pm$ 0.042	0.017 $\pm$ 0.013
Ball in Cup Catch	Visited States	Final	2655.000 $\pm$ 55.154	2680.500 $\pm$ 3.536	2731.500 $\pm$ 14.849	2664.500 $\pm$ 12.021
		AUC	2415.110 $\pm$ 30.751	2406.878 $\pm$ 69.883	2499.878 $\pm$ 14.418	2334.732 $\pm$ 19.627
	Episode Reward	Final	958.730 $\pm$ 2.447	939.715 $\pm$ 29.013	964.390 $\pm$ 1.909	945.705 $\pm$ 13.513
		AUC	708.565 $\pm$ 87.088	544.440 $\pm$ 262.849	716.214 $\pm$ 27.763	399.726 $\pm$ 85.061
Finger Spin	Visited States	Final	5541.500 $\pm$ 34.648	5428.500 $\pm$ 58.690	5412.000 $\pm$ 156.978	5513.500 $\pm$ 41.719
		AUC	5317.105 $\pm$ 4.489	5202.047 $\pm$ 88.931	5195.651 $\pm$ 213.415	5317.360 $\pm$ 34.418
	Episode Reward	Final	965.350 $\pm$ 6.604	957.740 $\pm$ 0.905	956.460 $\pm$ 3.281	927.550 $\pm$ 41.111
		AUC	877.415 $\pm$ 11.824	880.650 $\pm$ 12.565	881.395 $\pm$ 3.270	846.177 $\pm$ 28.606

both ranking second among the evaluated methods and remaining close to the corresponding RIDE values of 964.4 ± 1.9 and 716.2 ± 27.8.

On `finger_spin`, HTC^2 obtains the highest final coverage, 5541.5 ± 34.6, followed by E3B at 5513.5 ± 41.7. The coverage AUC values of HTC^2 and E3B are 5317.1 ± 4.5 and 5317.4 ± 34.4, respectively. HTC^2 also achieves the highest final reward, 965.4 ± 6.6. For reward AUC, RIDE and ICM obtain 881.4 ± 3.3 and 880.7 ± 12.6, while HTC^2 reaches 877.4 ± 11.8.

Across the three tasks, HTC^2 achieves the highest final visited-state count on `cartpole_swingup` and `finger_spin`, while RIDE achieves the highest final coverage on `ball_in_cup_catch`. The corresponding reward statistics show the same diversity across environments, with HTC^2 attaining the highest final reward on `cartpole_swingup` and `finger_spin`, and the second-highest final reward on `ball_in_cup_catch`.

C.3 Additional Hamiltonian Latent Diagnostics

This section provides additional diagnostics for the learned Hamiltonian phase representation. The main text shows that the learned scalar $H_\varphi(q, p)$ is stable under passive evolution, responds to policy actions, and varies smoothly over the projected q and p subspaces. Here, we further inspect the

internal organization of the phase latents from three complementary perspectives: temporal phase trajectories, action-conditioned phase updates, and dimension-wise q–p coupling.

Phase trajectory organization. Figure 8 visualizes the coordinate-like latent q and the momentum-like latent p using separate PCA bases. For each subspace, we show three views: coloring by the learned Hamiltonian energy, coloring by normalized episode time, and projected trajectory traces. Both q and p exhibit smooth energy variation along the projected trajectory, while the normalized-time coloring shows an ordered temporal progression. The trajectory traces further indicate that the learned phase state evolves along coherent low-dimensional paths. This supports the use of posterior phase trajectories for same-episode future pair sampling in HTC^2.

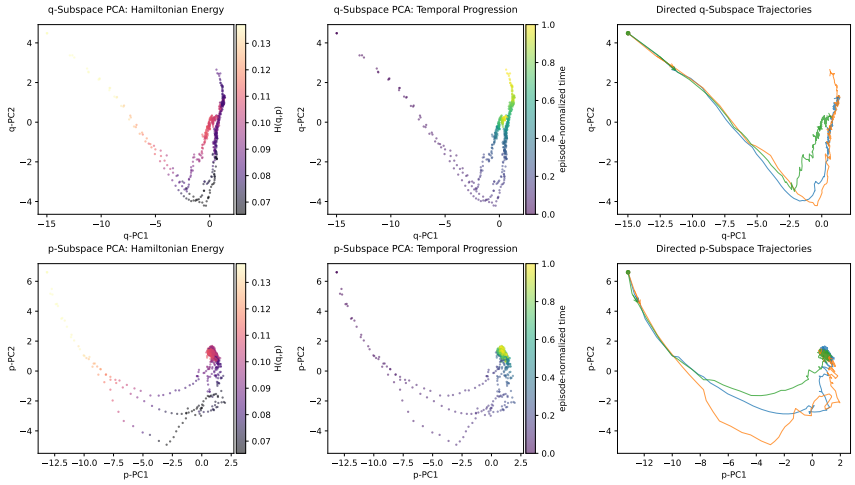


Figure 8: **Separate PCA projections of coordinate-like q and momentum-like p latents on `walker_walk`.** Each subspace is visualized by learned Hamiltonian energy $H_\varphi(q, p)$, normalized episode time, and projected trajectory traces. The learned phase representation exhibits smooth energy organization and temporally coherent evolution in both subspaces.

Action-conditioned phase evolution. Figure 9 analyzes how actions modulate the learned phase dynamics. The phase PCA colored by action magnitude and learned control push shows that control-related quantities vary systematically over the phase trajectory rather than being uniformly distributed. The one-step phase updates further separate low-control and high-control transitions: high-control samples produce larger and more directionally distinct phase updates. The phase-step magnitude statistics also increase for the high-control group. These results indicate that the learned phase state is responsive to action-conditioned control, which is the transition structure used by HTC^2 to construct horizon-conditioned future pairs.

Dimension-wise q–p coupling. Figure 10 reports the dimension-wise absolute correlation between the coordinate-like subspace q and the momentum-like subspace p on `walker_walk`, `hopper_hop`, and `cheetah_run`. Instead of treating q and p as two monolithic vectors, this analysis inspects all 16×16 cross-dimensional pairs. The heatmaps show that the q–p relation is not uniformly coupled across all dimensions: strong correlations appear only for selected dimension pairs, while many other pairs remain weakly or moderately correlated. This pattern indicates that the learned phase representation does not collapse q and p into a single mixed latent block. Rather, the two subspaces share the same trajectory-level dynamics while preserving a degree of dimension-wise separation.

This result is consistent with the intended phase-space structure of HTC^2. The coordinate-like and momentum-like components should not be fully independent, because they are inferred from the same controlled trajectory and jointly determine the learned Hamiltonian $H_\varphi(q, p)$. At the same time, they should not become identical or uniformly entangled. The observed local and dimension-selective coupling therefore supports the use of a structured phase representation for Hamiltonian temporal contrastive exploration.

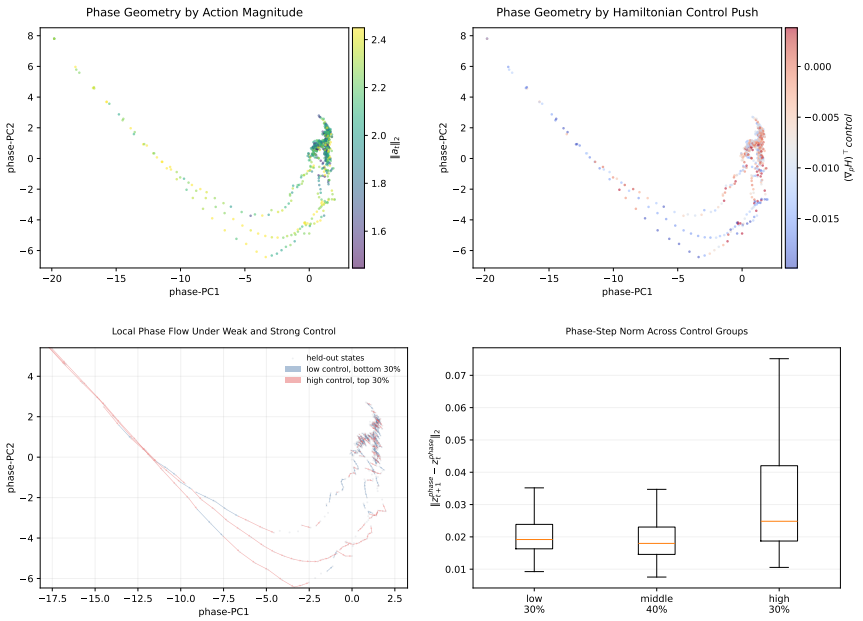


Figure 9: **Action-conditioned phase geometry on `walker_walk`.** The phase PCA is colored by action magnitude and learned control push. One-step phase updates are grouped by control strength, and the high-control group produces larger phase-step magnitudes. This shows that the learned phase dynamics are modulated by action-conditioned control.

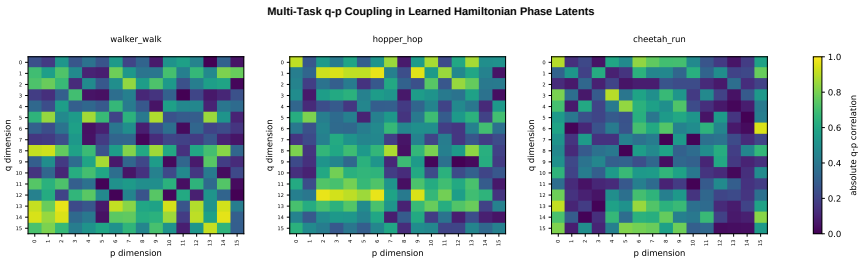


Figure 10: **Dimension-wise q–p coupling across DMC tasks.** Each heatmap shows the absolute correlation between the 16 dimensions of coordinate-like q and the 16 dimensions of momentum-like p. The correlations are local and dimension-selective rather than uniformly high, indicating that q and p remain dynamically related but do not collapse into a single fully mixed latent representation.

Summary. These diagnostics provide a finer view of the Hamiltonian phase representation used by HTC^2. The separate q/p PCA plots show energy-ordered and time-coherent phase trajectories; the action-conditioned analysis shows that phase updates are modulated by control strength; and the cross-task q–p coupling maps show trajectory-level dependence without uniform dimension-wise collapse. Together, these results support the representational structure assumed by HTC^2: exploration is driven by temporal future discovery in a Hamiltonian phase space whose components are dynamically related yet not reduced to a single entangled latent cloud.

C.4 Detailed Parametric OOD Results

This section reports the detailed numerical values for the parametric OOD policy-retention experiment in Section 4.5. Retention is computed as

$$\text{Retention} = 100 \times \frac{R_{\text{OOD}}}{R_{\text{ID}}},$$

where R_{OOD} is the zero-shot return under the perturbed environment and R_{ID} is the return in the original environment for the same method and seed. We evaluate four physical shifts: gravity ($1.5\times$), friction ($2.0\times$), mass ($1.5\times$), and damping ($2.0\times$).

Table 6: Parametric OOD zero-shot policy retention on `walker_walk`. Values report return retention relative to the base environment (%, mean ± std).

Model	Gravity 1.5×	Friction 2.0×	Mass 1.5×	Damping 2.0×
SAC+ICM	69.58 ± 3.02	94.10 ± 6.05	64.64 ± 13.19	101.45 ± 0.08
SAC+RIDE	58.79 ± 8.01	95.85 ± 0.21	62.80 ± 13.66	**101.73 ± 0.13**
SAC+E3B	70.08 ± 16.26	97.36 ± 1.56	70.65 ± 14.16	100.46 ± 1.79
SAC+HTC^2	**78.70 ± 11.94**	**100.01 ± 0.83**	**82.38 ± 10.40**	97.68 ± 3.05

Table 7: Parametric OOD zero-shot policy retention on `hopper_stand`. Values report return retention relative to the base environment (%, mean ± std).

Model	Gravity 1.5×	Friction 2.0×	Mass 1.5×	Damping 2.0×
SAC+ICM	5.42 ± 6.95	99.11 ± 1.21	40.76 ± 0.90	99.45 ± 1.37
SAC+RIDE	**52.37 ± 14.84**	96.55 ± 1.28	64.32 ± 15.83	**100.44 ± 0.30**
SAC+E3B	4.23 ± 3.48	94.85 ± 2.71	32.28 ± 17.39	99.60 ± 0.23
SAC+HTC^2	30.21 ± 5.63	**99.15 ± 1.36**	**85.73 ± 10.99**	97.65 ± 3.06

Discussion. The detailed results show that HTC^2 obtains the best retention on seven out of twelve OOD settings. The strongest pattern appears under mass perturbations, where HTC^2 achieves the highest retention on all three tasks. It also achieves the best gravity-shift retention on `walker_walk` and `cheetah_run`. These results are consistent with the objective of HTC^2: exploration is guided by multi-step temporal discovery in a Hamiltonian phase representation, which encourages policies to experience physically informative transitions during training. Friction and damping shifts show more task-dependent behavior, indicating that different intrinsic rewards can favor different robustness profiles.

C.5 Ablation Learning Curves

We further analyze the sample efficiency of the ablation variants in Figure 11. The curves compare full HTC^2 with variants that remove the physical consistency loss, remove the Hamiltonian energy gate, replace temporal contrastive reward with one-step ΔH, remove the RND (Burda et al., 2018b) warm start, or remove representation normalization.

On `walker_walk`, full HTC^2 improves steadily and reaches the highest final return. Several variants learn useful behaviors, but the full method maintains the strongest late-stage performance.

On `hopper_hop`, the difference between variants is more pronounced: full HTC^2 reaches high return and remains stable, whereas the one-step ΔH reward fails to learn an effective hopping behavior. The variant without representation normalization also suffers a large drop, indicating that stable contrastive distances are important for this underactuated task.

On `cheetah_run`, full HTC^2 again achieves the strongest final performance. The variants without $\mathcal{L}_{\text{phys}}$, without the energy gate, or without RND (Burda et al., 2018b) warm start improve during training but converge to lower returns. These learning curves are consistent with the final-return ablation in Table 1: the complete HTC^2 design improves both sample efficiency and final performance by combining Hamiltonian phase learning with multi-step temporal contrastive curiosity.

C.6 Policy Rollout Visualization

To qualitatively inspect the behavior induced by the final HTC^2 policy, we perform additional environment rollouts from the trained checkpoint and sample temporally ordered observations from six DMC tasks. Figure 12 shows twelve sampled frames ($t = 0, \ldots, 11$) for `walker_walk`, `walker_run`, `hopper_hop`, `hopper_stand`, `cheetah_run`, and `walker_stand`.

The sampled trajectories exhibit distinct task-dependent motion patterns. On `walker_walk`, the agent remains upright while its leg configuration evolves across the rollout, whereas `walker_run` exhibits substantially larger changes in body and limb configuration. The hopper remains close to an upright posture on `hopper_stand`, while `hopper_hop` shows greater temporal variation in its articulated configuration. On `cheetah_run`, the torso and limbs evolve through a sequence of locomotion configurations. For `walker_stand`, the rollout progresses from an initially inclined configuration toward a more upright posture, followed by continued balancing adjustments.

Table 8: Parametric OOD zero-shot policy retention on `cheetah_run`. Values report return retention relative to the base environment (%, mean ± std).

Model	Gravity 1.5×	Friction 2.0×	Mass 1.5×	Damping 2.0×
SAC+ICM	70.93 ± 2.46	54.84 ± 25.55	89.33 ± 17.98	43.74 ± 9.29
SAC+RIDE	77.72 ± 9.16	**92.31 ± 8.88**	100.90 ± 7.78	63.70 ± 4.68
SAC+E3B	74.47 ± 27.46	61.43 ± 7.62	96.43 ± 18.18	**67.63 ± 6.52**
SAC+HTC^2	**84.18 ± 2.15**	64.67 ± 38.44	**107.85 ± 8.09**	59.89 ± 4.71

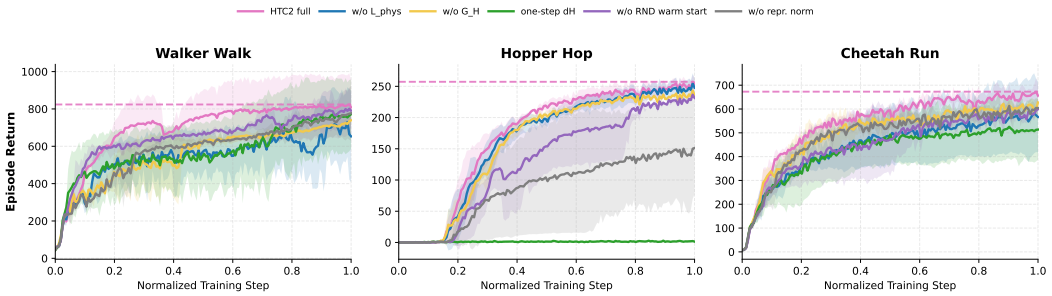


Figure 11: **Ablation learning curves.** Curves report mean performance over 4 seeds, and shaded regions denote standard deviation. Full HTC^2 is compared with variants removing the physical consistency loss, Hamiltonian energy gate, RND warm start, representation normalization, and with a one-step ΔH reward.



Figure 12: **HTC^2 policy rollouts on six DMC tasks.** Each row corresponds to one task, and columns show twelve temporally ordered sampled frames ($t = 0, \ldots, 11$) from the final trained policy checkpoint. The visualization is used only as a qualitative inspection of policy behavior.