HAMILTONIAN WORLD MODEL FOR PHYSICALLY CONSISTENT EMBODIED PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied world models are increasingly expected to serve not only as video generators, but as internal simulators for action-conditioned prediction. However, most existing models parameterize latent transitions with generative modules, leaving physical structure to be learned implicitly from data. This design may accumulate physical drift over long horizons, especially in contact-rich manipulation. We propose Hamiltonian World Model (`HWM`), a physically structured latent transition framework. The core concept of `HWM` is to move physical inductive bias from auxiliary regularization or post-hoc evaluation into the transition law itself: backbone latents are mapped into a learned phase space of generalized coordinates and momenta, evolved through dissipative Hamiltonian dynamics. A bidirectional phase-space interface and a multi-step transition-consistency objective keep the structured rollout compatible with existing world-model encoders and decoders, making `HWM` a drop-in replacement for the latent transition module. In simulation, `HWM` improves speed, acceleration, and jerk prediction while better preserving system energy than unstructured transitions. We further instantiate `HWM` across three embodied world-model settings: Ctrl-World for single-arm multi-view manipulation, EVAC for dual-arm single-view manipulation, and Cosmos-Predict2.5-2B for single-arm single-view task execution. On Robocasa, Cosmos-Predict2.5-2B with `HWM` achieves relative success-rate gains of 2.7%; on real-robot AgiBot World data, `HWM` delivers relative gains of 9.7% on nDTW with Ctrl-World and 8.4% on Logic with EVAC. These results suggest that physically structuring the latent transition is a practical path toward embodied world models whose imagined futures are more stable, more physically faithful, and more useful for action. The project page is available at `https://hamilton-world-model.github.io/Hamiltonian_World_Model/`.

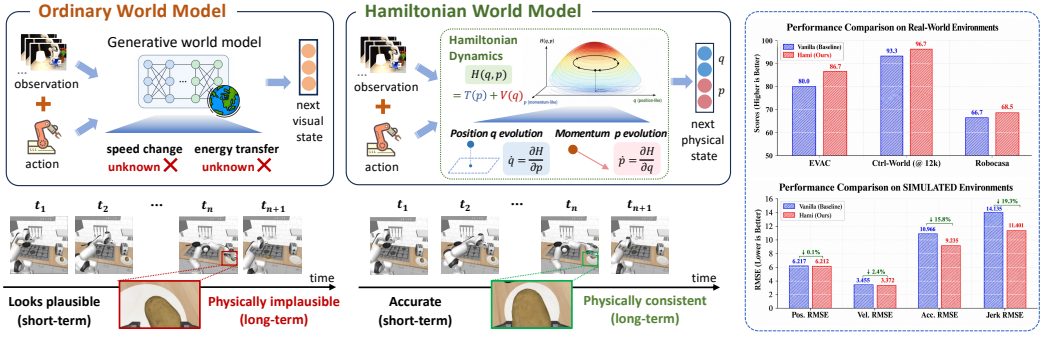

Figure 1: Comparison between ordinary generative world models and our Hamiltonian World Model.

1 INTRODUCTION

World models have long been studied as internal simulators for sequential decision making, enabling planning, counterfactual reasoning, and data-efficient policy improvement (Sutton, 1991; Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2023). Recent advances in generative video modeling have pushed this paradigm into embodied manipulation, where a model is expected to predict action-conditioned futures for robotic interaction (Hu et al., 2023; Bruce et al., 2024; Seo et al., 2023; Zhou

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.

et al., 2024; Zhu et al., 2025; Guo et al., 2025; NVIDIA et al., 2025). In this setting, however, the value of a world model is not determined only by visual realism. A predicted future must also remain physically coherent enough to support downstream policy evaluation and long-horizon execution.

This requirement exposes a central limitation of many modern embodied world models. Although current architectures can generate visually plausible short-term videos, their latent transitions are often implemented as generic sequence or generative modules. Such transitions can fit local appearance statistics and short-range correlations, but they do not explicitly encode the dynamical structure that governs physical interaction. As a result, small one-step errors may compound into long-horizon physical drift: objects may penetrate, contact timing may become inconsistent, and action effects may gradually decouple from the commanded motion. For an embodied agent, these failures directly weaken the usefulness of imagined rollouts as substrates for planning and control.

Existing approaches address this issue from two directions. One line of work scales generative world models and improves controllability, view consistency, and open-loop prediction quality (Hu et al., 2023; Bruce et al., 2024; Zhou et al., 2024; Zhu et al., 2025; Guo et al., 2025; NVIDIA et al., 2025). These models provide strong perceptual and rendering capabilities, but the physical regularities of manipulation are still mostly learned implicitlym, shown in Figure 3. Another line introduces physical structure into learned dynamics. For example, Symplectic ODE-Nets show that structure-preserving formulations can improve stability and extrapolation in dynamical systems (Greydanus et al., 2019; Toth et al., 2020; Zhong et al., 2020). Yet the embodied video world-model setting requires a different design target: a transition mechanism that can be inserted into modern latent visual world models while preserving their existing observation, tokenization, and rendering interfaces.

Our key conceptual claim is that physical consistency should not be treated as an auxiliary property imposed after a generic transition has been learned. Instead, the transition law itself should be physically structured. In other words, the central modeling object should shift from "a neural sequence model that predicts the next latent" to "a learned physical evolution rule in latent phase space." This view is natural for embodied manipulation: a robot does not merely observe image sequences, but interacts with a dynamical system through forces, contacts, inertia, and dissipation. A world model that aims to support action should represent not only appearance evolution, but also the latent physical state whose evolution determines whether an imagined future remains executable.

Motivated by this view, we propose Hamiltonian World Model (`HWM`), a physically structured latent transition framework for embodied world modeling. `HWM` maps backbone latents into learned generalized coordinates and momenta, which evolve under a dissipative Hamiltonian system. The Hamiltonian models the energy landscape, while damping accounts for non-conservative effects such as friction. A semi-implicit Euler update reduces numerical drift, and the predicted phase states are mapped back to backbone latents for decoding. This design structures long-horizon dynamics while preserving the backbone's perceptual and rendering capabilities.

The framework offers three conceptual advantages. First, coordinates and momenta separate configuration-like information from motion-related dynamics. Second, an energy-based transition combines Hamiltonian evolution with learned dissipation. Third, multi-step latent consistency aligns the structured rollout with backbone trajectories over horizons where physical drift emerges. Together, these components form a plug-and-play transition module compatible with existing observation interfaces.

We evaluate `HWM` in simulation and three embodied world-model settings. In simulation, it improves speed, acceleration, and jerk prediction while better preserving energy relative to an unstructured transition. On Ctrl-World for single-arm multi-view manipulation and EVAC for dual-arm single-view manipulation, it improves generation and logic metrics, supporting better visual-temporal and physical consistency. On Cosmos-Predict2.5-2B for single-arm single-view task execution, it increases success rates on LIBERO and Robocasa, demonstrating benefits for both prediction quality and downstream execution.

Our contributions are summarized as follows:

1. We propose a transition-centric view of world modeling: physical consistency should be built into the latent transition law, rather than added through auxiliary losses or post-hoc evaluation.

2. We introduce HWM, a plug-and-play Hamiltonian latent transition framework that maps backbone latents into phase-space variables and evolves them with dissipative Hamiltonian dynamics for improved long-horizon stability.
3. We demonstrate consistent benefits across simulation and three embodied world-model settings, including physical-state prediction, better visual and logic metrics on Ctrl-World and EVAC, and higher downstream task success on LIBERO and Robocasa with Cosmos. The anonymous code of our framework is accessible at `https://anonymous.4open.science/r/iclr-anonymous-hami-5CA4`.

2 Related Work

Generative and embodied video world models. Recent work has scaled world models into high-capacity video generators and interactive simulators. GAIA-1 and Genie demonstrated that generative models can support controllable, action-conditioned future prediction at scale (Hu et al., 2023; Bruce et al., 2024). In robotics, manipulation-oriented models such as Multi-View Masked World Models, RoboDreamer, IRASim, Ctrl-World, and Cosmos further improved view consistency, controllability, and open-loop rollout quality for embodied interaction (Seo et al., 2023; Zhou et al., 2024; Zhu et al., 2025; Guo et al., 2025; NVIDIA et al., 2025). Broader embodied-data and system efforts, such as AgiBot World, likewise reflect the field's shift toward scalable, manipulation-centered world modeling (AgiBot-World-Contributors et al., 2025). Our work is complementary to this line: rather than proposing a new rendering or tokenization backbone, HWM replaces the generic latent transition with a physically structured alternative. This distinction matters because current systems can achieve strong perceptual realism while still learning dynamics mostly implicitly, which leaves room for physical drift over long rollouts.

Physics-structured neural dynamics and world models. Another line of work introduces explicit physical structure into learned dynamics. Hamiltonian Neural Networks, Hamiltonian Generative Networks, and Symplectic ODE-Nets showed that structure-preserving parameterizations can improve stability and extrapolation in dynamical systems (Greydanus et al., 2019; Toth et al., 2020; Zhong et al., 2020). More recent world-model variants, including HaM-World, PH-Dreamer, and LaWM, inject Hamiltonian, port-Hamiltonian, or least-action principles into latent transitions (Tang et al., 2026; Luan & Shi, 2026; Xiao & Ghaffari, 2026). HWM is most closely related to this family, but differs in emphasis: we aim for a drop-in transition design for embodied video world models that jointly captures conservative evolution, control, dissipation, and residual contact-rich effects, instead of focusing only on planner-oriented control benchmarks or on a single physical formalism.

3 Preliminaries

We briefly review two building blocks of our approach: learned world models for robotic systems, and Hamiltonian dynamics as an inductive bias for structured state transitions.

3.1 World Models for Embodied Agents

We learn a forward model in a low-dimensional latent space rather than directly from raw observations. An encoder e_ϕ maps an observation o_t to a latent state z_t, a dynamics predictor f_θ advances z_t to z_{t+1} given action a_t, and a decoder d_ψ reconstructs or predicts future observations $\hat{o}_t$. The model is trained by minimizing the multi-step prediction error, e.g.,

$$\mathcal{L} = \mathbb{E}_{(o_t,a_t)\sim\mathcal{D}} \sum_{t=1}^{T} \|o_t - d_\psi(z_t)\|^2 + \lambda \cdot \mathcal{R}(z_t), \tag{1}$$

where $\mathcal{R}(z_t)$ is a latent regularizer (e.g., KL divergence to a prior in variational settings).

Standard black-box dynamics trained this way tend to accumulate error and violate physical invariants over long rollouts, especially in contact-rich manipulation. To address this, we constrain f_θ to follow a controlled dissipative Hamiltonian structure in latent space, where z_t is split into generalized coordinates q_t and momenta p_t, and the transition is governed by a learnable energy function $H_\theta(q, p)$. This inductive bias enforces energy conservation (up to explicit dissipation and actuation) and symplectic geometry, which we formalize next.

3.2 HAMILTONIAN DYNAMICS

To formalize the physical structure used by HWM, we review Hamiltonian dynamics in a form suitable for latent-space world modeling. Unlike classical settings where q, p are physical coordinates, in our framework they are *learned latent phase variables* inferred from raw observations. Let

$$z_t = \begin{bmatrix} q_t \\ p_t \end{bmatrix} \in \mathbb{R}^{2d} \tag{2}$$

denote a latent phase state, where $q_t \in \mathbb{R}^d$ represents generalized latent coordinates and $p_t \in \mathbb{R}^d$ represents generalized latent momenta. A Hamiltonian $H_\theta(q, p)$ (parameterized by a neural network) defines the total energy of the latent system. In the separable case, it is often written as

$$H_\theta(q, p) = T_\theta(p) + V_\theta(q), \tag{3}$$

where $T_\theta(p)$ is the learned kinetic energy and $V_\theta(q)$ is the learned potential energy.

For a conservative latent system, the dynamics are governed by the canonical Hamilton equations

$$\dot{q} = \frac{\partial H_\theta}{\partial p}, \qquad \dot{p} = -\frac{\partial H_\theta}{\partial q}. \tag{4}$$

Equivalently, by defining the symplectic matrix

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}, \tag{5}$$

the latent dynamics can be written compactly as

$$\dot{z} = J \nabla_z H_\theta(z), \tag{6}$$

where $\nabla_z H_\theta(z)$ denotes the energy gradient in latent phase space. Since $J^\top = -J$, Eq. equation 6 preserves the Hamiltonian along trajectories of an autonomous conservative latent system:

$$\frac{d}{dt} H_\theta(z_t) = \nabla_z H_\theta(z_t)^\top \dot{z}_t = \nabla_z H_\theta(z_t)^\top J \nabla_z H_\theta(z_t) = 0. \tag{7}$$

Embodied manipulation, however, is not purely conservative: robot actions inject energy, while friction, damping, and contact dissipation remove it. To model this setting, we consider a *controlled dissipative Hamiltonian* system in the latent space:

$$\dot{z} = \big(J - R_\phi(z)\big) \nabla_z H_\theta(z) + B_\psi(z) u + r_\omega(z, u), \tag{8}$$

where $R_\phi(z) \succeq 0$ is a learned dissipation matrix, u is the robot control input, $B_\psi(z)$ maps controls into latent phase-space forces, and $r_\omega(z, u)$ captures residual non-conservative effects such as contact impulses, deformation, or modeling error. The corresponding energy change in the latent space is

$$\frac{d}{dt} H_\theta(z_t) = -\nabla_z H_\theta(z_t)^\top R_\phi(z_t) \nabla_z H_\theta(z_t) + \nabla_z H_\theta(z_t)^\top B_\psi(z_t) u_t + \nabla_z H_\theta(z_t)^\top r_\omega(z_t, u_t). \tag{9}$$

Equation equation 9 makes explicit how conservative transport, dissipation, actuation, and residual contact effects contribute differently to the evolution of the latent energy, providing a strong inductive bias for long-horizon prediction.

For long-horizon prediction, the discretization of Eq. equation 8 is as important as the continuous-time dynamics. A naive explicit Euler update can accumulate substantial geometric drift, whereas structure-preserving integrators better retain the phase-space organization of the underlying system (Zhong et al., 2020). A standard symplectic Euler discretization of the conservative core is

$$p_{t+1} = p_t - \Delta t \, \nabla_q H_\theta(q_t, p_{t+1}), \qquad q_{t+1} = q_t + \Delta t \, \nabla_p H_\theta(q_t, p_{t+1}), \tag{10}$$

which can be extended with control, dissipation, and residual terms. In HWM, observations are mapped to latent phase variables (q_t, p_t) via e_ϕ, the dominant transition is evolved by the controlled dissipative Hamiltonian rule, and additional residual channels r_ω absorb non-ideal effects that are unavoidable in real manipulation.

4 HAMILTONIAN LATENT WORLD MODELS FOR EMBODIED AGENTS

We present HWM, a Hamiltonian latent world model for embodied agents. Our framework follows the standard world-model architecture: an encoder maps observations to latent states, a structured dynamics module advances the latent state over time, and a decoder renders predictions back to observation space. The key departure from prior work lies in the dynamics module: instead of relying solely on unconstrained neural latent transitions, we introduce a Hamiltonian phase-space evolution that enforces physical inductive biases in the latent space.

4.1 WHY HAMILTONIAN DYNAMICS?

World models learn latent dynamics that approximate the evolution of observed physical systems. A variational perspective provides a structured way to describe such dynamics through stationary paths of an action functional $\mathcal{S}[q] = \int_{t_0}^{t_1} L(q, \dot{q}; t)\, dt$, where L is a Lagrangian. These paths satisfy the Euler–Lagrange equations $\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} = \frac{\partial L}{\partial q}$.

Hamiltonian mechanics reformulates this principle in phase space. Introducing generalized momenta $p = \partial L/\partial \dot{q}$, an invertible Legendre transform yields the Hamiltonian $H(q, p; t) = p^\top \dot{q} - L(q, \dot{q}; t)$ and Hamilton's equations $\dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q}$. This reformulation preserves the variational dynamics while making their symplectic structure explicit.

For latent world models, this connection motivates Hamiltonian dynamics as a structured inductive bias rather than a necessary consequence of prediction-error minimization. A learned Hamiltonian $H_\theta(q, p)$ constrains latent evolution through phase-space geometry while allowing the energy landscape to be learned from data. For robotic manipulation, we augment this conservative core with action-dependent forcing and learned dissipation $D_\theta(p)$ to account for external inputs, friction, and damping. The resulting controlled dissipative dynamics retain a symplectic conservative component while accommodating non-conservative effects.

4.2 LATENT PHASE-SPACE ENCODING

As illustrated in Fig. 2, the backbone encoder E_ϕ first maps the observation o_t to a latent representation

$$z_t = E_\phi(o_t), \qquad z_t \in \mathbb{R}^{d_z}. \tag{11}$$

To introduce a structured dynamical representation, HWM further maps the latent state z_t into a Hamiltonian phase space consisting of generalized coordinates and momenta: $(q_t, p_t) = \phi_\theta(z_t)$, where $q_t, p_t \in \mathbb{R}^d$ denote latent generalized coordinates and momenta, respectively.

This phase-space representation serves two purposes. First, it explicitly introduces the coordinate–momentum structure required by Hamiltonian mechanics, where physical dynamics are represented as a first-order system in (q, p). Second, the explicit momentum variable provides an independent representation of velocity- and inertia-related information, which is important for modeling systems with non-negligible momentum and temporal persistence.

To maintain compatibility with the original backbone latent space, we introduce a learnable inverse mapping $\hat{z}_t = \psi_\theta(q_t, p_t)$, which maps a phase-space state back to the latent representation used by the backbone decoder. Therefore, ϕ_θ and ψ_θ form a bidirectional interface between the original latent space and the structured Hamiltonian phase space.

We parameterize the latent Hamiltonian using a separable kinetic–potential form:

$$H_\theta(q_t, p_t) = \frac{1}{2} p_t^\top M_\theta^{-1}(q_t) p_t + V_\theta(q_t), \tag{12}$$

where $M_\theta^{-1}(q_t)$ denotes a learned inverse mass matrix and $V_\theta(q_t)$ denotes the learned potential energy. The first term corresponds to kinetic energy, while the second captures potential energy.

The separable structure $T(q, p) + V(q)$ provides a physically motivated inductive bias while retaining the representational flexibility of neural networks. In particular, the quadratic dependence on momentum makes the momentum contribution to the kinetic energy explicit, while the configuration-dependent mass matrix $M_\theta(q_t)$ allows the effective inertia of the latent system to vary with its configuration. This is particularly suitable for articulated or highly nonlinear systems whose inertial properties depend on their state.

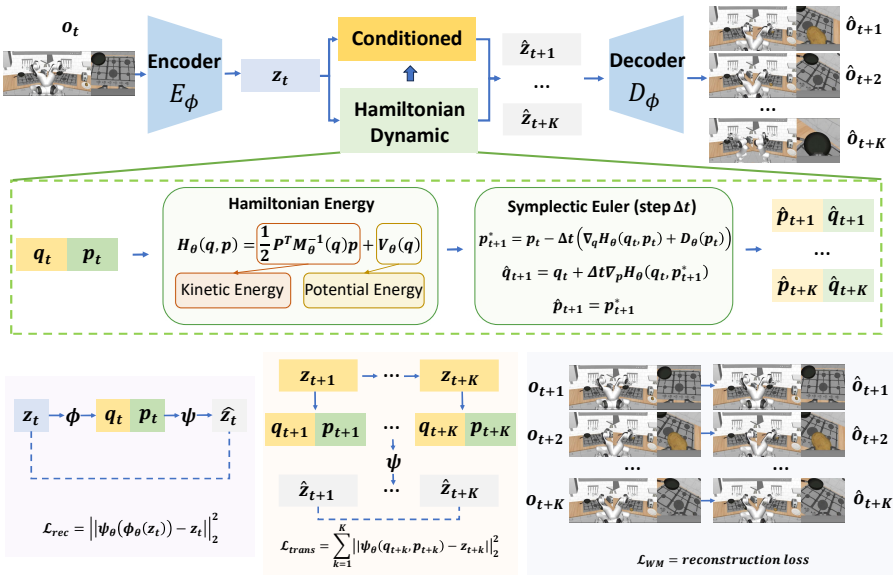

Figure 2: Overview of the proposed Hamiltonian latent dynamics framework. Given an observation o_t, the encoder E_ϕ maps it to a latent state z_t, which is decomposed into generalized coordinates and momenta, (q_t, p_t), and evolved over multiple future steps using a learned hamiltonian energy function and a Symplectic Euler integrator. The resulting latent trajectory is mapped to latent representations and decoded by D_ϕ to predict future observations. Training is jointly guided by the latent reconstruction loss, latent transition loss, and world-model reconstruction loss.

4.3 Hamiltonian Dynamics in Latent Space

Given the Hamiltonian in Eq. equation 12, the latent dynamics are governed by a dissipative Hamiltonian system:

$$\dot{q}_t = \nabla_p H_\theta(q_t, p_t), \tag{13}$$

$$\dot{p}_t = -\nabla_q H_\theta(q_t, p_t) - D_\theta(p_t), \tag{14}$$

where $D_\theta(p_t)$ denotes a learned damping term.

The first terms in Eqs. equation 13–equation 14 describe the conservative Hamiltonian dynamics, while $D_\theta(p_t)$ accounts for dissipative effects such as friction and unmodeled energy loss. Following the formulation illustrated in Fig. 2, we use a momentum-dependent damping function of the form $D_\theta(p_t) = \gamma_\theta(p_t)\, p_t$, where $\gamma_\theta(\cdot)$ is a learned scalar-valued function.

This choice provides a simple and effective dissipation mechanism while preserving the Hamiltonian structure of the conservative component. Compared with a fully parameterized dissipation matrix, the momentum-proportional formulation introduces fewer additional degrees of freedom and reduces the complexity of learning, while remaining expressive enough to capture the dominant energy-loss effects in the target dynamics.

4.4 Structure-Preserving Rollout

To predict future latent states, we discretize the continuous dynamics using the semi-implicit symplectic Euler scheme shown in Fig. 2. Given the current phase-space state (q_t, p_t), one rollout step is defined as

$$p^*_{t+1} = p_t - \Delta t\left(\nabla_q H_\theta(q_t, p_t) + D_\theta(p_t)\right), \tag{15}$$

$$\hat{q}_{t+1} = q_t + \Delta t\, \nabla_p H_\theta(q_t, p^*_{t+1}), \tag{16}$$

$$\hat{p}_{t+1} = p^*_{t+1}. \tag{17}$$

The updated phase-space state is therefore $(\hat{q}_{t+1}, \hat{p}_{t+1})$, which is recursively used to generate a K-step prediction trajectory: $\{(\hat{q}_{t+k}, \hat{p}_{t+k})\}_{k=1}^{K}$.

The key feature of the semi-implicit symplectic Euler scheme is that the position update evaluates $\nabla_p H_\theta$ using the newly updated momentum p^*_{t+1}. Compared with a naive explicit Euler discretization,

this update better preserves the geometric structure of the conservative Hamiltonian system and reduces numerical energy drift over long-horizon rollouts. The damping term is evaluated explicitly, which provides a practical compromise between numerical structure preservation and computational efficiency.

After the phase-space rollout, each predicted state is projected back into the original backbone latent space through the inverse phase-space mapping:

$$\hat{z}_{t+k} = \psi_\theta \left(\hat{q}_{t+k}, \hat{p}_{t+k} \right), \qquad k = 1, \ldots, K. \tag{18}$$

The corresponding future observations are then obtained using the backbone decoder:

$$\hat{o}_{t+k} = D_\phi(\hat{z}_{t+k}), \qquad k = 1, \ldots, K. \tag{19}$$

Thus, as illustrated in Fig. 2, the complete prediction process consists of three stages: (i) encoding the current observation into the backbone latent space, (ii) evolving the latent state in the structured Hamiltonian phase space, and (iii) mapping the predicted phase states back to the latent and observation spaces.

4.5 Training Objective

The Hamiltonian dynamics module is trained jointly with the backbone world model using the following multi-objective loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{WM}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}}, \tag{20}$$

where $\mathcal{L}*\text{WM}$ denotes the observation-level reconstruction loss, $\mathcal{L}*\text{rec}$ enforces consistency between the latent and phase-space representations, and $\mathcal{L}_{\text{trans}}$ supervises the multi-step Hamiltonian rollout.

First, the phase-space reconstruction loss is defined as $\mathcal{L}_{\text{rec}} = \left\| \psi_\theta \left(\phi_\theta(z_t) \right) - z_t \right\|_2^2$. This objective encourages ϕ_θ and ψ_θ to behave as approximately inverse mappings. It ensures that the learned phase-space representation remains compatible with the backbone latent space and prevents the model from learning a dynamically convenient but semantically disconnected latent coordinate system.

Second, we introduce a transition consistency loss over the entire prediction horizon:

$$\mathcal{L}_{\text{trans}} = \sum_{k=1}^{K} \left\| \psi_\theta \left(\hat{q}_{t+k}, \hat{p}_{t+k} \right) - z_{t+k} \right\|_2^2, \tag{21}$$

where the target future latent state is obtained from the ground-truth observation:

$$z_{t+k} = E_\phi(o_{t+k}), \qquad k = 1, \ldots, K. \tag{22}$$

The transition consistency loss directly constrains the Hamiltonian rollout to remain on the latent manifold induced by the backbone encoder. In contrast to one-step supervision, the multi-step formulation penalizes accumulated prediction error and discourages the dynamics model from drifting into phase-space regions that are dynamically plausible under the learned Hamiltonian but inconsistent with the actual latent trajectories.

Finally, the observation-level world-model loss is computed between the predicted and ground-truth future observations: $\mathcal{L}_{\text{WM}} = \sum_{k=1}^{K} \ell_{\text{WM}} \left(\hat{o}_{t+k}, o_{t+k} \right)$, where ℓ_{WM} denotes the reconstruction loss adopted by the backbone world model.

Together, these three losses establish consistency across the phase-space, latent-space, and observation-space levels. Specifically, $\mathcal{L}_{\text{rec}}$ constrains the phase-space representation, $\mathcal{L}_{\text{trans}}$ constrains the temporal evolution of the structured latent dynamics, and $\mathcal{L}_{\text{WM}}$ ensures that the resulting predictions remain accurate in observation space. All components are differentiable and optimized jointly through backpropagation through the multi-step Hamiltonian rollout.

5 Experiments

We evaluate `HWM` from three complementary perspectives: physical-state prediction and energy consistency in simulation, visual generation quality, and long-horizon physical consistency and

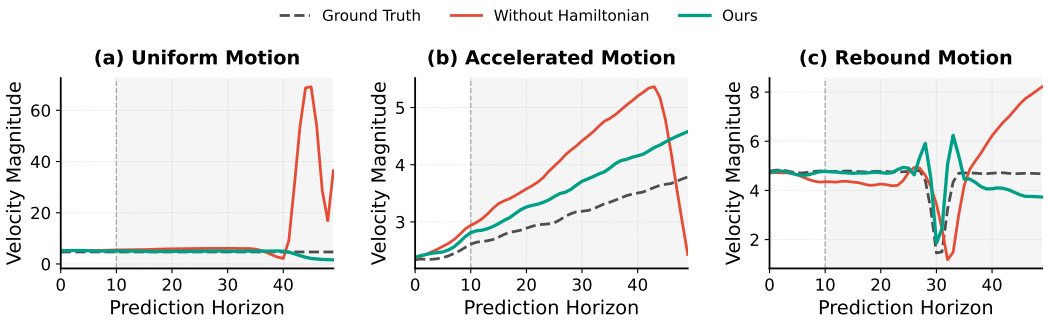

Figure 3: Qualitative comparison of the predicted velocity curves.

downstream task success in real-world rollouts. In simulation, we test whether HWM captures underlying physical regularities beyond short-range correlations. For real-world evaluation, we consider three world-model families: COSMOS for single-arm single-view prediction (NVIDIA et al., 2025), Ctrl-World for single-arm multi-view manipulation (Guo et al., 2025), and EVAC (Jiang et al., 2025) for dual-arm single-view manipulation, using each backbone's standard visual generation metrics. We emphasize Ctrl-World for long-horizon analysis, as multi-view observations reveal temporal and geometric inconsistencies. Across experiments, we preserve the original observation interface and other backbone components, replacing only the latent transition with HWM where applicable and comparing primarily against the corresponding vanilla backbone.

5.1 Simulation Experiments

To isolate the effect of latent transition modeling from high-dimensional visual rendering, we first evaluate HWM in simulation using SimVP as the baseline, where accurate physical annotations enable direct assessment of learned dynamics. Detailed simulation settings are provided in Appendix B.

Instantaneous Physical State Prediction. We evaluate whether HWM better captures instantaneous physical states, including position, speed, acceleration, and jerk. As shown in Table 1, HWM consistently reduces prediction errors across all metrics and rollout horizons. The predicted velocity trajectories in Fig. 3 also remain closer to the ground truth, indicating that Hamiltonian structuring improves the accuracy and stability of the learned dynamics.

Table 1: Instantaneous physical-state prediction under different rollout horizons. All models are trained with a 10-step prediction horizon. Lower values indicate better performance.

Method	Step	Pos. RMSE	Vel. RMSE	Acc. RMSE	Jerk RMSE
Vanilla	30	6.217	3.455	10.966	14.135
HWM	30	**6.212**	**3.372**	**9.235**	**11.401**
Vanilla	50	5.590	2.545	8.332	12.329
HWM	50	**5.524**	**2.525**	**7.483**	**10.750**

Energy Conservation. We further evaluate long-horizon physical consistency through Hamiltonian energy drift. As shown in Fig. 4, HWM maintains energy trajectories substantially closer to the ground truth, while the non-Hamiltonian variant exhibits larger drift and oscillations. This demonstrates that the proposed structured dynamics better preserve global physical consistency over long rollouts.

Table 2: Quantitative comparison between HWM and Vanilla on Ctrl-World.

Method	nDTW	PSNR	BLEU	Scene (%)	CLIP (%)	Logic (%)
Vanilla	0.670	10.13	0.767	90.88	78.31	93.3
HWM	**0.735**	**10.16**	**0.802**	**90.96**	**78.96**	**96.7**

5.2 Real-World Large World-Model Experiments: Visual Generation Quality

We evaluate HWM as a drop-in transition module in Ctrl-World (Guo et al., 2025) and EVAC (Jiang et al., 2025). As shown in Table 2, nDTW increases from 0.670 to 0.735, BLEU from 0.767 to 0.802, and Logic from 93.3% to 96.7% on Ctrl-World. On EVAC, Table 3 shows that BLEU increases from 0.104 to 0.111 and Logic from 80.0% to 86.7%, with improvements in the remaining reported metrics. These results support improved multi-view visual consistency and generation coherence for bimanual

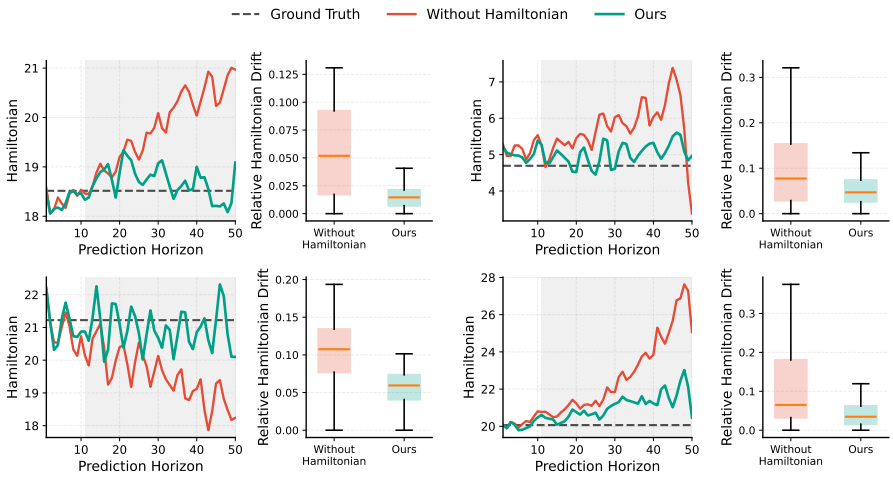

Figure 4: Comparison of energy conservation across different methods.

Table 3: Comparison between HWM and Vanilla models. Metrics are collected from two evaluation settings on EVAC.

Method	nDTW	PSNR	BLEU	Scene (%)	CLIP (%)	Logic (%)
Vanilla	0.282	16.61	0.104	91.3	77.5	80.0
HWM	**0.283**	**16.64**	**0.111**	**92.2**	**78.4**	**86.7**

Table 4: Overall success rate comparison on LIBERO and Robocasa benchmarks.

Method	LIBERO					Robocasa
	Spatial (%)	Object (%)	Goal (%)	Long (%)	Avg (%)	Avg (%)
Vanilla	97.6	99.8	96.2	97.6	97.8	66.7
HWM	**98.6**	**99.8**	**98.0**	**98.2**	**98.7**	**68.5**

interactions across different architectures and embodiment settings. Detailed settings and evaluation protocols appear in Appendix C.

5.3 Real-World Large World-Model Experiments: Task Execution Success Rate

We integrate HWM into *Cosmos-Predict2.5-2B* and evaluate downstream task execution on LIBERO (Liu et al., 2023) and RoboCasa (Nasiriany et al., 2024). As shown in Table 4, average success increases from 97.8% to 98.7% on LIBERO, with matching or improved performance across all four suites, and from 66.7% to 68.5% on RoboCasa. These results support the usefulness of Hamiltonian transitions for downstream embodied execution. Evaluation protocols, task-level results across seeds, and qualitative examples appear in Appendix D.

6 Conclusion

We presented HWM, a Hamiltonian transition framework that combines controlled dissipative dynamics with a residual branch to improve long-horizon physical consistency in embodied world models. Experiments across Ctrl-World, COSMOS, and EVAC demonstrate benefits in prediction quality, physical consistency, and downstream task execution. Several limitations remain. Due to limited computational resources, we could not evaluate more backbones or conduct broader experiments across tasks, training scales, and random seeds. Moreover, learned phase variables need not correspond to identifiable physical quantities, so Hamiltonian structure alone does not guarantee physically correct predictions. Dissipation, external control, and the residual branch also limit the conservation properties of the complete dynamics, particularly during contact-rich interactions. Finally, the current evaluations do not establish robustness to substantially different environments or rollout horizons beyond those tested. Broader evaluation and stronger characterization of the learned dynamics are therefore needed.

REFERENCES

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL https://arxiv.org/abs/2503.06669.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Luis Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation, 2025. URL https://arxiv.org/abs/2510.10125.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, and Ruihai Wu. Robowm-bench: A benchmark for evaluating world models in robotic manipulation, 2026. URL https://arxiv.org/abs/2604.19092.

Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Xueyu Luan and Chenwei Shi. Ph-dreamer: A physics-driven world model via port-hamiltonian generative dynamics, 2026. URL https://arxiv.org/abs/2605.18303.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation, 2023. URL https://arxiv.org/abs/2302.02408.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.

Haoyun Tang, Haodong Cui, Keyao Xu, Kun Wang, and Zhandong Mei. Ham-world: Soft-hamiltonian world models with selective memory for planning, 2026. URL https://arxiv.org/abs/2605.05951.

Peter Toth, Danilo J. Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. In *International Conference on Learning Representations*, 2020.

Ruixiang Wang, Qingming Liu, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. Eva: Aligning video world models with executable robot actions via inverse dynamics rewards, 2026. URL https://arxiv.org/abs/2603.17808.

Qixin Xiao and Maani Ghaffari. Lawm: Least action world models for long-horizon physical consistency from visual observations. *arXiv preprint arXiv:2605.08279*, 2026.

Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. In *International Conference on Learning Representations*, 2020.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination, 2024. URL https://arxiv.org/abs/2404.12377.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2406.14540.

Table 5: Quantitative comparison of video prediction under different training and evaluation horizons. Lower values indicate better performance.

Train Horizon	Test Horizon	MSE ↓		MAE ↓	
		Vanilla	HWM	Vanilla	HWM
10	10	2.76	**2.56**	23.88	**21.27**
10	30	25.38	**23.82**	53.56	**49.27**
10	50	55.57	**51.18**	87.59	**80.37**
10	60	62.32	**53.03**	95.32	**82.58**
30	30	25.81	**24.38**	79.36	**73.31**
30	50	37.55	**33.65**	94.54	**87.84**
30	60	44.59	**38.93**	102.55	**94.60**

A Related Work

Latent world models for planning and control. World models have long served as learned simulators for sequential decision making, from Dyna and World Models to PlaNet, Dreamer, and DreamerV3 (Sutton, 1991; Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2023). These works established that compact latent dynamics can support imagination-based planning and control, but they were primarily developed for reinforcement learning settings with comparatively simple observation and interaction structure. HWM builds on this planner-facing latent-transition perspective, but targets embodied manipulation scenarios in which viewpoint changes, contact events, and action-conditioned rollouts place much stronger demands on long-horizon physical consistency.

Embodiment-grounded evaluation of world models. As world models are increasingly used for robot learning, evaluation has moved beyond appearance metrics toward action-conditioned usefulness and executability. Recent studies argue that generated futures should support policy ranking or downstream control rather than merely look plausible (Guo et al., 2025; Wang et al., 2026). RoboWM-Bench makes this embodiment-grounded perspective explicit by evaluating whether generated manipulation videos can be translated into successful robot actions (Jiang et al., 2026). Our evaluation protocol follows this shift: we assess HWM not only by predictive quality, but by whether its structured transitions improve long-horizon physical consistency and action usefulness across distinct embodied world-model families.

B Simulation Experiments

To separate transition dynamics modeling from high-dimensional rendering complexity, we first conduct controlled experiments in a simulated environment. Because the simulator provides clean action-conditioned trajectories together with accurate physical annotations, it enables a more direct test of whether the learned latent dynamics preserve the underlying physical structure. Under this setting, we adopt SimVP as the baseline and investigate the effectiveness of incorporating the proposed Hamiltonian World Model (HWM) . We evaluate HWM from three complementary perspectives: long-horizon prediction quality, instantaneous physical-state prediction, and energy conservation.

Long-Horizon Prediction Quality. We first investigate whether incorporating Hamiltonian structure improves the overall quality of predicted trajectories. Specifically, we train the models with prediction horizons of 10 and 30 frames, and roll them out to different test horizons to evaluate both in-distribution prediction and long-horizon extrapolation. The quantitative results are reported in Table 5. Across all combinations of training and test horizons, HWM consistently achieves lower MSE and MAE than the vanilla baseline, showing that imposing Hamiltonian structure also improves visual prediction quality. More importantly, the advantage becomes more pronounced under long-horizon rollout. For example, when trained with a 10-frame prediction horizon and extrapolated to 60 frames, HWM reduces the MSE from 62.32 to 53.03 and the MAE from 95.32 to 82.58, corresponding to relative reductions of 14.9% and 13.4%, respectively. These results suggest that Hamiltonian modeling mitigates the accumulation of dynamics errors during rollout, leading to more stable and accurate long-term predictions.

Instantaneous Physical State Prediction. Beyond overall prediction quality, we further evaluate the accuracy of instantaneous physical-state prediction, including object position, speed, acceleration, and jerk. For each generated rollout, we first estimate the object center in every frame and compute velocity using finite differences between consecutive positions. Higher-order physical quantities, including acceleration and jerk, are then derived accordingly. Our evaluation includes the full HWM, an ablated variant without the Hamiltonian constraint, and the ground-truth trajectory. We fix the training prediction horizon to 10 frames and roll out the models to 30 and 50 frames. The results are summarized in Table 1. HWM consistently yields lower physical-state prediction errors under both rollout horizons. Notably, the improvements are substantially larger for higher-order motion quantities. Under the 30-frame rollout, HWM reduces the acceleration RMSE from 10.966 to 9.235, corresponding to a 15.8% reduction, and decreases the jerk RMSE from 14.135 to 11.401, yielding a 19.3% improvement. The same trend persists under the longer 50-frame rollout, where the acceleration and jerk errors are reduced by 10.2% and 12.8%, respectively. These results demonstrate that Hamiltonian modeling improves the ability to capture instantaneous system dynamics and recover accurate motion states. We further visualize the predicted velocity trajectories in Fig. 3. The velocity curves produced by HWM exhibit substantially better agreement with the ground truth, whereas the non-Hamiltonian variant shows larger deviations and stronger fluctuations, further validating the effectiveness of the Hamiltonian constraint for instantaneous physical-state prediction.

Energy Conservation. Finally, we evaluate the physical consistency of generated trajectories from the perspective of energy conservation. For each rollout, we estimate the underlying physical states from the generated video and compute the corresponding Hamiltonian(total system energy). We then compare the temporal evolution of the Hamiltonian produced by HWM, the non-Hamiltonian variant, and the ground truth. For conservative systems, such as projectile motion under uniform gravity, the total mechanical energy should remain approximately constant. Therefore, Hamiltonian drift provides a principled measure of long-horizon physical consistency. We further visualize both the Hamiltonian trajectories and the corresponding energy drift in Fig. 4. As shown in the figure, the Hamiltonian predicted by HWM remains significantly closer to the ground-truth energy throughout the rollout, whereas the non-Hamiltonian variant exhibits substantially larger drift and oscillations. These results demonstrate that the proposed Hamiltonian formulation effectively preserves global physical invariants, leading to long-horizon dynamics that are more faithful to the underlying physical system.

C Real-World Large World-Model Experiments: Visual Generation Quality

We next evaluate HWM as a drop-in transition module in real large world models. The key question here is whether a more physically grounded latent transition can improve not only physical plausibility, but also the standard visual generation quality used in practice to assess modern embodied world models.

Ctrl-World. On Ctrl-World, which emphasizes controllable single-arm multi-view manipulation (Guo et al., 2025), HWM improves visual quality while simultaneously strengthening cross-view consistency. Replacing the generic transition reduces view-dependent drift and better preserves the correspondence between commanded actions and the motion observed from different cameras, yielding more stable open-loop video generation across viewpoints.

EVAC. On EVAC (Jiang et al., 2025), which targets dual-arm single-view manipulation, HWM again improves visual generation quality over the original backbone. The gain is most apparent in coordinated bimanual interactions, where physically implausible drift in relative motion and contact timing is otherwise amplified over time. This result shows that the benefit of Hamiltonian structuring is not tied to a single observation layout or embodiment regime, but transfers across substantially different world-model families.

Details. We train our model on the large-scale real-robot dataset *AgiBot World Challenge*, which encompasses 10 typical task categories across domestic, industrial, and medical scenarios. For evaluation, we adopt *EWMBench* (Yue et al., 2025) as the benchmark, which quantifies generation quality from five dimensions: appearance fidelity (PSNR, SSIM), temporal scene consistency (Scene), gripper trajectory alignment (nDTW), semantic description similarity (BLEU, CLIP), and task/physical logic (Logic). All fidelity metrics are higher-the-better. The test set of EWMBench is derived

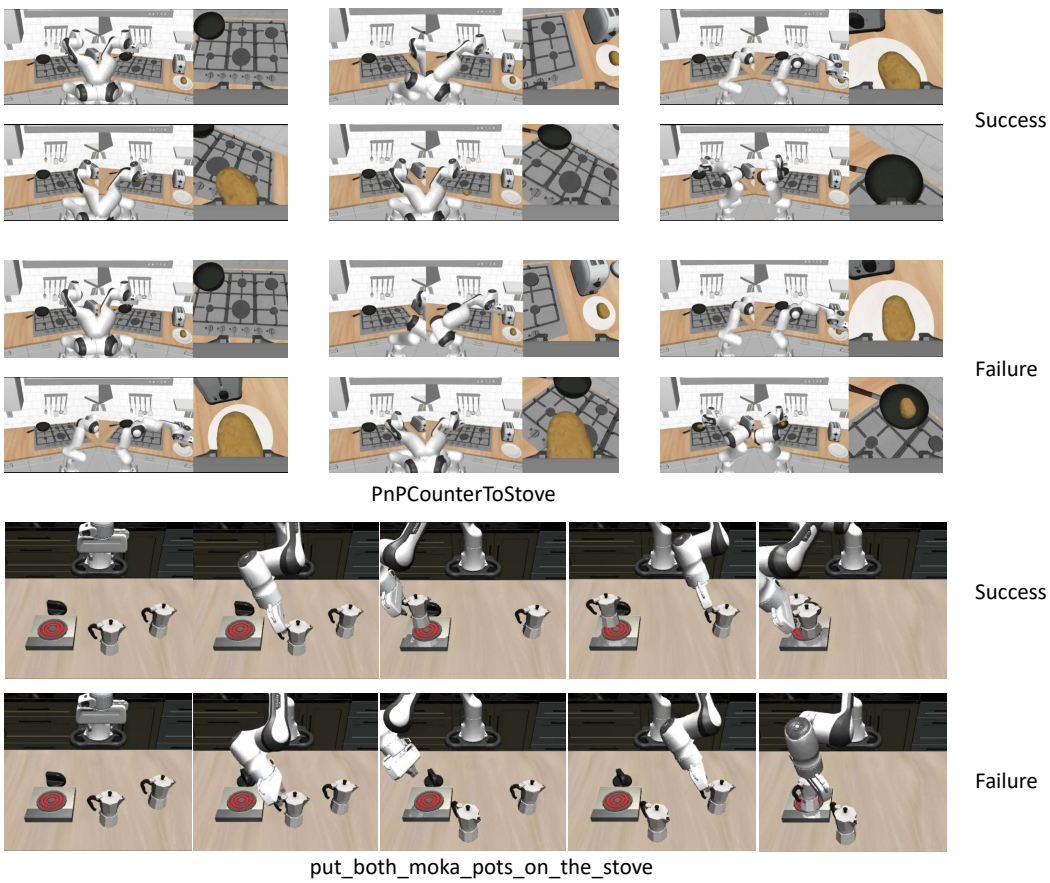

Figure 5: Examples of HWM (successful) and Vanilla (failed) executions on two representative tasks: PnP (Pick-and-Place) and Put. The top rows show the PnP task (PnPCounterToStove), while the bottom rows show the Put task (put_both_moka_pots_on_the_stove). For each task, successful and failed trajectories are presented for comparison.

from held-out samples of AgiBot World, and the evaluation pipeline is fully automated to ensure reproducible results.

D Real-World Large World-Model Experiments: Task Execution Success Rate

COSMOS. To evaluate whether the proposed Hamiltonian latent dynamics improve the practical usefulness of large world models for embodied decision making, we conduct task execution experiments with *Cosmos-Predict2.5-2B* as the backbone world model. Rather than evaluating visual prediction quality in isolation, we measure whether imagined future trajectories remain sufficiently consistent with the underlying environment dynamics to support successful task execution. We evaluate HWM on two widely used simulation benchmarks, LIBERO and Robocasa, which cover both long-horizon manipulation and diverse kitchen interaction tasks. As shown in Table 4, HWM consistently outperforms the vanilla model on both benchmarks. On LIBERO, our method achieves an average success rate of 98.7%, improving over the vanilla model by 0.9 percentage points, with gains across all four task categories and the largest improvement on the Goal subset (from 96.2% to 98.0%). On the more challenging Robocasa benchmark, HWM further improves the average success rate from 66.7% to 68.5%. These results indicate that the proposed Hamiltonian latent dynamics produce more reliable imagined trajectories, resulting in more effective downstream task execution. To further understand the source of these improvements, we provide qualitative case studies in Figure 5, which compare successful executions of HWM against failure cases of the vanilla model on two representative manipulation tasks.

Table 6: Task-level success rates across three seeds on COSMOS with Robocasa. V and H denote Vanilla and HWM, respectively.

Task	V195	V196	V197	V_{avg}	H195	H196	H197	H_{avg}	Δ
CloseDoubleDoor	88	84	88	86.7	90	86	96	90.7	+4.0
CloseDrawer	98	98	98	98.0	98	100	100	99.3	+1.3
CloseSingleDoor	96	96	96	96.0	96	92	100	96.0	0
CoffeePressButton	98	96	98	97.3	96	94	100	96.7	-0.7
CoffeeServeMug	58	52	48	52.7	56	58	62	58.7	+6.0
CoffeeSetupMug	40	20	38	32.7	38	28	44	36.7	+4.0
OpenDoubleDoor	84	92	90	88.7	80	88	86	84.7	-4.0
OpenDrawer	92	96	76	88.0	94	96	100	96.7	+8.7
OpenSingleDoor	80	88	90	86.0	82	90	88	86.7	+0.7
PnPCabToCounter	28	38	30	32.0	30	30	36	32.0	0
PnPCounterToCab	62	58	50	56.7	60	58	66	61.3	+4.7
PnPCounterToMicrowave	26	40	26	30.7	30	36	36	34.0	+3.3
PnPCounterToSink	62	50	52	54.7	58	48	64	56.7	+2.0
PnPCounterToStove	50	74	66	63.3	50	68	56	58.0	-5.3
PnPMicrowaveToCounter	40	34	24	32.7	40	46	46	44.0	+11.3
PnPSinkToCounter	74	70	70	71.3	70	76	76	74.0	+2.7
PnPStoveToCounter	68	72	60	66.7	60	66	66	64.0	-2.7
TurnOffMicrowave	100	100	100	100	100	100	100	100	0
TurnOffSinkFaucet	70	76	78	74.7	72	78	78	76.0	+1.3
TurnOffStove	20	24	28	24.0	14	20	20	18.0	-6.0
TurnOnMicrowave	76	76	84	78.7	76	82	82	80.0	+1.3
TurnOnSinkFaucet	64	60	60	61.3	64	70	70	68.0	+6.7
TurnOnStove	52	54	50	52.0	52	58	58	56.0	+4.0
TurnSinkSpout	72	80	76	76.0	72	78	78	76.0	0

LIBERO. We follow the standard LIBERO simulation benchmark protocol (Liu et al., 2023). LIBERO consists of a single Franka Emika Panda robot arm interacting with a variety of objects and environments, and provides four primary task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. These suites evaluate different aspects of manipulation generalization, including spatial layout variation, object variation, language-specified goals, and long-horizon task composition. Each suite contains 10 tasks with 50 demonstrations per task, resulting in 500 demonstrations per task suite.

For our experiments, Cosmos-Predict2.5-2B is used as the base world model, and we compare the original transition model against HWM while keeping the remaining policy and planning components fixed. Following the standard data processing protocol (Kim et al., 2026), unsuccessful demonstrations are filtered from the policy-training data, while the complete unfiltered demonstration set is retained for world-model and value-function training. We report task success rate across the four LIBERO suites, with particular emphasis on LIBERO-Long, where accumulated transition errors over extended action sequences provide a more demanding test of the learned dynamics.

The results show that incorporating Hamiltonian latent dynamics consistently improves task execution success over the original Cosmos transition. The gains are especially pronounced on long-horizon tasks, where small one-step prediction errors can otherwise compound into increasingly inaccurate imagined trajectories. By enforcing a structured evolution in latent phase space, HWM produces more stable intermediate states and more reliable action-conditioned predictions, allowing the downstream policy to maintain a coherent sequence of actions for a longer duration. These results indicate that improvements in latent transition quality translate directly into better downstream manipulation performance rather than merely improving reconstruction fidelity.

Robocasa. We further evaluate HWM on the Robocasa simulation benchmark (Nasiriany et al., 2024), which contains 24 static kitchen manipulation tasks executed by a single Franka Emika Panda robot arm. Following the evaluation protocol of prior works (Kim et al., 2026), we evaluate each task over 50 trials distributed across five evaluation scenes with different floor plans and visual styles, with 10 trials per scene. The final success rate is averaged across all 24 tasks and three random seeds, corresponding to 3600 evaluation trials in total.

Unlike LIBERO, the Robocasa evaluation protocol emphasizes generalization to unseen object instances, and two of the five evaluation scenes for each task contain styles that are not observed during training. This setting therefore provides a stronger test of whether the learned world model captures transferable interaction dynamics rather than memorizing appearance-specific correlations.

To isolate the data efficiency of the proposed dynamics model, we train on the 50 human-teleoperated demonstrations available for each task and do not use the additional MimicGen demonstrations. As in the LIBERO experiments, unsuccessful demonstrations are removed for policy training, while the complete demonstration set is retained for world-model and value-function training. We then evaluate the task execution success rate under the standard Robocasa protocol.

Compared with the original Cosmos transition model, `HWM` achieves higher average task success across the 24 Robocasa tasks. The improvement is particularly meaningful in scenes containing unseen styles and object instances, where purely appearance-driven prediction can be insufficient to maintain consistent object interactions over long action sequences. The structured Hamiltonian transition reduces accumulated latent drift and produces more coherent action-conditioned futures, which in turn improves the reliability of downstream planning and policy execution.

Overall, the LIBERO and Robocasa results demonstrate that the benefit of `HWM` extends beyond improved latent or visual prediction metrics. When integrated into Cosmos-Predict2.5-2B, the proposed Hamiltonian transition consistently improves downstream task execution success, with the largest gains appearing in settings that require long-horizon reasoning, sustained object interaction, and generalization beyond the visual configurations encountered during training. These results provide direct evidence that physically structured latent dynamics can improve the usefulness of large world models for embodied control.

E AI USE STATEMENT

We used large language models (LLMs) to assist with language polishing, including improving grammar, readability, and presentation of the manuscript. LLMs were not used to generate experimental results or make scientific conclusions. All technical content, mathematical formulations, experimental designs, analyses, and reported results were developed and verified by the authors. The authors take full responsibility for the content of the paper.

F REPRODUCIBILITY STATEMENT

We provide an anonymized code repository containing the implementation of the proposed Hamiltonian World Model (HWM) to support reproducibility. The paper and appendix describe the model architecture, latent phase-space parameterization, Hamiltonian dynamics, numerical integration scheme, training objectives, experimental settings, evaluation protocols, and benchmark configurations used in our experiments.

We report the relevant training and evaluation horizons, baseline configurations, benchmark protocols, and random-seed evaluations where applicable. Our experiments cover both controlled simulation and multiple embodied world-model backbones, with the corresponding experimental details provided in the appendix. Together with the anonymized implementation, these details are intended to enable reproduction of the main results reported in the paper.

G ETHICS STATEMENT

This work does not involve human subjects, sensitive personal data, or deployment on physical robots that could directly endanger users. Our experiments focus on improving the reliability of learned world models for embodied interaction. More physically consistent world models could help reduce unsafe action proposals by making model-based evaluation more faithful, but they could also strengthen autonomous systems in ways that require responsible downstream use. We therefore emphasize transparent evaluation, careful reporting, and reproducible methodology.

To promote reproducibility and transparency, we provide detailed experimental settings, algorithmic descriptions, and anonymized code. Our contribution is methodological and aims to improve the

reliability of embodied world modeling rather than bypass safety constraints. We are committed to the principles outlined in the ICLR Code of Ethics and confirm that this work adheres to them.