Structured Dynamics Optimization: Expert-Calibrated Energy Modulation of Multi-Motion Predictive Priors for Humanoid Imitation

Ange Tong
Independent Researcher, China
Corresponding author: 2024140315@stu.hit.edu.cn

Abstract

Frame-wise imitation provides a clear reference but not a measure of whether a learned future deserves equal influence at every state. We present **Structured Dynamics Optimization (SDO)**, a reward-level mechanism that converts internal discrepancies of a pretrained motion model into weak, state-dependent guidance for reinforcement learning. The high-level model is trained first on a multi-motion corpus containing dozens of expert trajectories, learning shared temporal structure before low-level policy optimization begins. From three past 67-dimensional states, a Transformer produces a 256-dimensional latent code and a 30-frame coarse future; a 20-step diffusion-inspired refiner produces the corrected future. SDO reads two quantities from this frozen model: expert-calibrated latent deviation and coarse-to-refined correction. Their calibrated sum forms an energy proxy that controls how strongly the predicted target is trusted, while a separate policy–prior agreement term checks its physical relevance to the current state. The auxiliary term enters only through the reward interface. In LocoMuJoCo, SDO energy is positively associated with future prediction error, and refinement energy reaches a Spearman correlation of about 0.55 at 15- and 30-step horizons. Across five paired 300-million-step runs, SDO raises deterministic imitation return from 16.4396 to 16.9191 (+2.92%) and episode length from 125.75 to 134.61; every paired seed improves on both measures. Evaluation disables the SDO bonus for both methods, so the gains reflect the policy learned during training rather than an altered test-time reward.

Keywords: humanoid control; imitation learning; motion prediction; motion prior; reinforcement learning; energy-based modulation

1 Introduction

A learned motion prior can carry information that a frame-wise imitation reward does not: where a movement is heading, which future states are compatible with the recent history, and how the current motion fits into a broader repertoire. That additional structure is attractive in physics-based control, where DeepMimic and later humanoid systems have already shown that reinforcement learning can track demanding reference motion while respecting body dynamics [1, 4, 6]. The field has consequently moved from literal tracking toward distributional priors, reusable skill spaces, structured latent dynamics, and generative planners. The unresolved interface problem considered here is more specific. Once the reference is supplemented by a learned predictor, the controller needs a way to distinguish a well-supported forecast from one that required the predictor to extrapolate or heavily revise its own estimate.

This matters especially when the predictor is not trained for a single motion. The high-level model used here is pretrained on a corpus containing dozens of expert trajectories drawn from multiple motion collections. Its purpose is to learn reusable temporal structure before the low-level policy sees the formal control task. A three-frame history is embedded into a latent representation and decoded into a 30-frame future, then passed through an iterative refinement stage. Such a model carries more than a target state: it exposes where the history lies in the learned representation and how much the forecast changes while being refined. Those internal quantities are available before the policy is rewarded for following the prediction.

Structured Dynamics Optimization (SDO) turns that internal evidence into a control signal. The name reflects the path by which the signal is constructed. *Structure* comes from the multi-motion latent representation learned from expert trajectories. *Dynamics* comes from the coarse future and its time-conditioned refinement.

Optimization is the final use of this information as a small gain on an existing reinforcement-learning objective rather than as a replacement optimizer. In the implemented form, the predicted target provides a direction in physical state space, while energy controls the authority of that direction.

The distinction from ordinary confidence-weighted tracking is upstream of the final reward formula. SDO does not estimate a generic uncertainty score from policy error. It reads two different failure cues from two different stages of the pretrained predictor: latent deviation from fixed expert statistics and the magnitude of the full-horizon coarse-to-refined correction. Each cue is calibrated on expert motion, the two are combined in energy space, and only then is the result converted to a reliability gain. Policy–prior agreement is computed separately. A prediction can therefore be close to the current policy state yet receive little influence because the predictor itself regards the underlying history or future construction as atypical; conversely, a low-energy prediction does not generate a large bonus if the physical state is incompatible with it.

The evidence is organized around that mechanism. Predictor-derived energy ranks future prediction difficulty from one to thirty steps, with the refinement term especially informative at medium and long horizons. A short diagnostic isolates the role of gating: the same predictive target becomes less useful when applied uniformly, whereas energy modulation restores the common evaluator. In the formal control study, five independently seeded 300M-step SDO runs all finish above their paired PPO baselines in deterministic imitation return, and all five also produce longer episodes. Test-time evaluation removes the SDO term entirely. Together, these results show that information already present inside a multi-motion predictor can be converted into a useful, non-dominant bias on policy learning without changing PPO itself.

2 Related work

2.1 From reference tracking to adaptive imitation

Physics-based imitation first demonstrated that difficult whole-body skills could be learned by combining reinforcement learning with explicit motion tracking. DeepMimic is the canonical example: carefully shaped pose, velocity, end-effector, and root objectives make high-fidelity tracking possible even for acrobatic behaviors [1]. Its strength is precise reference following, but the reference and the imitation objective remain tightly coupled; when motion libraries grow, choosing and weighting the right target becomes part of the control problem rather than a solved preprocessing step.

Later systems attacked that scaling problem from different directions. PHC showed that a single humanoid controller can absorb very large motion collections while learning recovery from failure states, shifting the emphasis from one-clip fidelity to persistent control over heterogeneous data [4]. MaskedMimic goes further on the interface side: by treating control as masked motion inpainting, one policy can respond to partial keyframes, scene constraints, or other incomplete motion specifications instead of requiring a separate reward design for each control modality [6]. These methods broaden what a reference-conditioned controller can track, but they still assume that the motion information presented to the controller is the information that should be used.

A second line of work changes the reference itself when that assumption breaks. Bi-Level Motion Imitation explicitly optimizes the motion target together with the policy so that human demonstrations can be reconciled with robot morphology and actuation limits [9]. MaskAdapt instead learns a mask-invariant base controller and a residual adaptation mechanism, preserving untargeted behavior while modifying selected body parts [10]. The important lesson for SDO is not that reference tracking is inadequate, but that modern imitation increasingly treats the reference–controller interface as a decision variable. SDO addresses a narrower version of that interface problem: when the reference is supplemented by a learned future prediction, it asks how much authority a particular prediction should receive before it is used as an auxiliary training signal.

2.2 Distributional, latent, and reusable motion priors

Motion-prior methods reduce dependence on literal frame-wise tracking by moving motion knowledge into a learned distribution or representation. AMP replaces hand-designed style tracking with an adversarial motion-quality reward learned from unstructured clips [2]. This is powerful because the controller can select and compose motions without an explicit clip scheduler. The discriminator score, however, answers whether generated behavior resembles the demonstration distribution; it is not designed to say whether one particular future produced by a predictive model is internally well supported.

ASE and PULSE push the same idea toward reusable skill spaces. ASE combines adversarial imitation with unsupervised skill discovery so that a large motion library becomes a controllable latent repertoire for downstream tasks [3]. PULSE expands the coverage substantially by distilling a universal humanoid representation from a large motion imitator and conditioning its prior on proprioception [5]. Their main design question is therefore *which behavior can be represented or selected through the latent space*. SDO leaves the low-level action space and policy representation unchanged; its question begins one step later, after a predictive prior has proposed a direction: *how strongly should this particular proposal influence learning?*

Structured latent dynamics provide an even closer comparison. FLD identifies periodic or quasi-periodic motion structure in a continuously parameterized Fourier latent space, enabling interpolation, online tracking, and a fallback strategy when risky targets are proposed [7]. This demonstrates that latent structure can support both motion generation and safety-aware control. SDO uses structure differently. Its latent representation is not the policy command space, and its refinement signal is not a fallback trigger. Instead, latent atypicality and coarse-to-refined disagreement are treated as continuous evidence that attenuates the gain of a predicted target. H-GAP represents another branch: it learns a generative state–action trajectory model and uses that model directly as a generalist planner under model-predictive control [8]. In that setting the generative model becomes the planning substrate; in SDO the pretrained model remains outside the low-level optimizer and contributes only a bounded reward-level modulation.

Recent work makes the distinction sharper. Score-Matching Motion Priors (SMP) convert a frozen motion diffusion model into a reusable task-agnostic reward through score distillation, avoiding the need to retrain an adversarial prior for every controller [11]. Context-Aware Motion Priors (CMP) then address a different failure mode of general priors: reference motions can be relevant to one task context and conflicting in another, so CMP learns context–motion compatibility and reweights prior supervision accordingly [12]. These developments show that "having a broad prior" and "using it uniformly" are separate design choices. SMP primarily asks how a generative model can provide a reusable motion-quality reward, while CMP asks which portions of a motion prior are relevant to the current task. SDO is complementary to both: it assumes a concrete predictive target has already been generated and estimates its *prediction-specific authority* from information internal to the predictor itself.

2.3 Future prediction and iterative generative refinement

Sequence models add another source of information that distributional priors do not necessarily expose: the trajectory by which a future prediction is constructed. Transformers are well suited to mapping short histories to multi-step futures because attention can couple distant coordinates and decode the entire horizon jointly [14]. Diffusion-based human-motion models such as MDM demonstrate that iterative denoising can represent multimodal motion sequences, while PhysDiff incorporates physics guidance to improve physical plausibility at generation time [16, 17]. In these works, the refinement process is mainly a means to obtain a better final sample.

The high-level model used by SDO is deliberately more modest. Its second stage borrows time-conditioned iterative reverse updates from diffusion models but is trained directly on future-sequence reconstruction. The relevant distinction is methodological: SDO does not claim a diffusion likelihood or a score-based uncertainty estimate. Instead, it observes a quantity that ordinary generation pipelines usually discard—the full-horizon displacement between the Transformer forecast and the refined future. That quantity is only a candidate difficulty signal until validated. The medium- and long-horizon error correlations reported later are therefore part of the method argument, not an after-the-fact interpretation of a diffusion score.

2.4 Uncertainty, out-of-distribution scores, and operational reliability

Predictive uncertainty can be estimated explicitly with Bayesian approximations or deep ensembles [18, 19]. These approaches are useful when calibrated predictive distributions are themselves required, but they introduce additional stochastic inference, multiple models, or an uncertainty head whose training objective is separate from the downstream controller. Representation-space alternatives are lighter. Mahalanobis-style feature distances and energy-based out-of-distribution scores use the geometry or energy of a learned representation to rank atypical inputs [20, 21].

SDO follows the latter operational philosophy but uses signals specific to motion prediction. Its latent term is a diagonal, expert-calibrated structural-deviation score in the temporal representation; its refinement term

measures disagreement between two stages of the same predictor. Neither quantity is presented as epistemic uncertainty, aleatoric uncertainty, or normalized likelihood. What SDO needs is a stable ordering: predictions that are structurally atypical or require unusually large internal correction should, on average, receive less authority if those signals track future error. Fixed expert quantiles make the two scores comparable, and the validation study tests the ordering directly.

Taken together, prior work has progressively improved four different capabilities: accurate tracking, reusable motion representations, generative or structured motion dynamics, and adaptation of a prior to downstream context. The remaining interface considered here is more local. Once a frozen multi-motion predictor has produced a particular future, its internal evidence is normally discarded and the target is either used or ignored as a whole. SDO turns that evidence into a continuous gain, separating *where the prior suggests the controller should move* from *how much that suggestion should be trusted*. This is the specific gap addressed by the method below.

3 Method

3.1 SDO as structured dynamics modulation

Let the trajectory-aligned humanoid state be

$$x_t = \begin{bmatrix} q_t \\ \dot{q}_t \end{bmatrix} \in \mathbb{R}^{67}, \tag{1}$$

with 34 generalized-position and 33 generalized-velocity components. Before low-level policy optimization, the high-level model is pretrained on a pooled multi-motion expert corpus containing dozens of valid trajectories. Every trajectory contributes normalized 67-dimensional history/future windows to the same representation-learning problem, with three past frames paired to a 30-frame future. The result is a shared predictive representation learned across a broader motion repertoire than the downstream control reference.

For a normalized history

$$H_t = [\tilde{x}_{t-N}, \ldots, \tilde{x}_{t-1}], \qquad N = 3, \tag{2}$$

the frozen high-level model produces

$$(C_t, z_t) = f_\theta(H_t), \qquad \hat{X}_t = g_\phi(C_t, z_t), \tag{3}$$

where $C_t \in \mathbb{R}^{30\times 67}$ is a coarse future, $z_t \in \mathbb{R}^{256}$ is the temporal latent code, and $\hat{X}_t \in \mathbb{R}^{30\times 67}$ is the refined future. SDO does not use these outputs as an unconditional extra target. It asks what the predictor itself reveals about the quality of the proposed future.

Two internal discrepancies answer that question. Latent energy measures how far z_t lies from fixed expert-calibrated latent statistics. Refinement energy measures how much the full predicted future moves between C_t and $\hat{X}_t$. Their calibrated combination is

$$E_{\mathrm{SDO}}(t) = E_{\mathrm{lat}}(t) + \lambda_{\mathrm{ref}} E_{\mathrm{ref}}(t). \tag{4}$$

The target and the energy play different roles. $\hat{X}_t$ supplies a candidate motion direction; E_{SDO} determines how much influence that direction receives. The implementation therefore realizes the original energy–manifold motivation without differentiating an explicit $-\nabla E$ through PPO. Structure and dynamics are learned in the high-level model; their discrepancy is exposed to the low-level learner as a bounded scalar gain.

Figure 1 summarizes the resulting interface. Multi-motion pretraining learns the reusable high-level representation once. The frozen model then constructs reference-history-conditioned priors offline, while the low-level reward path uses only the stored prior entry and current simulated state.

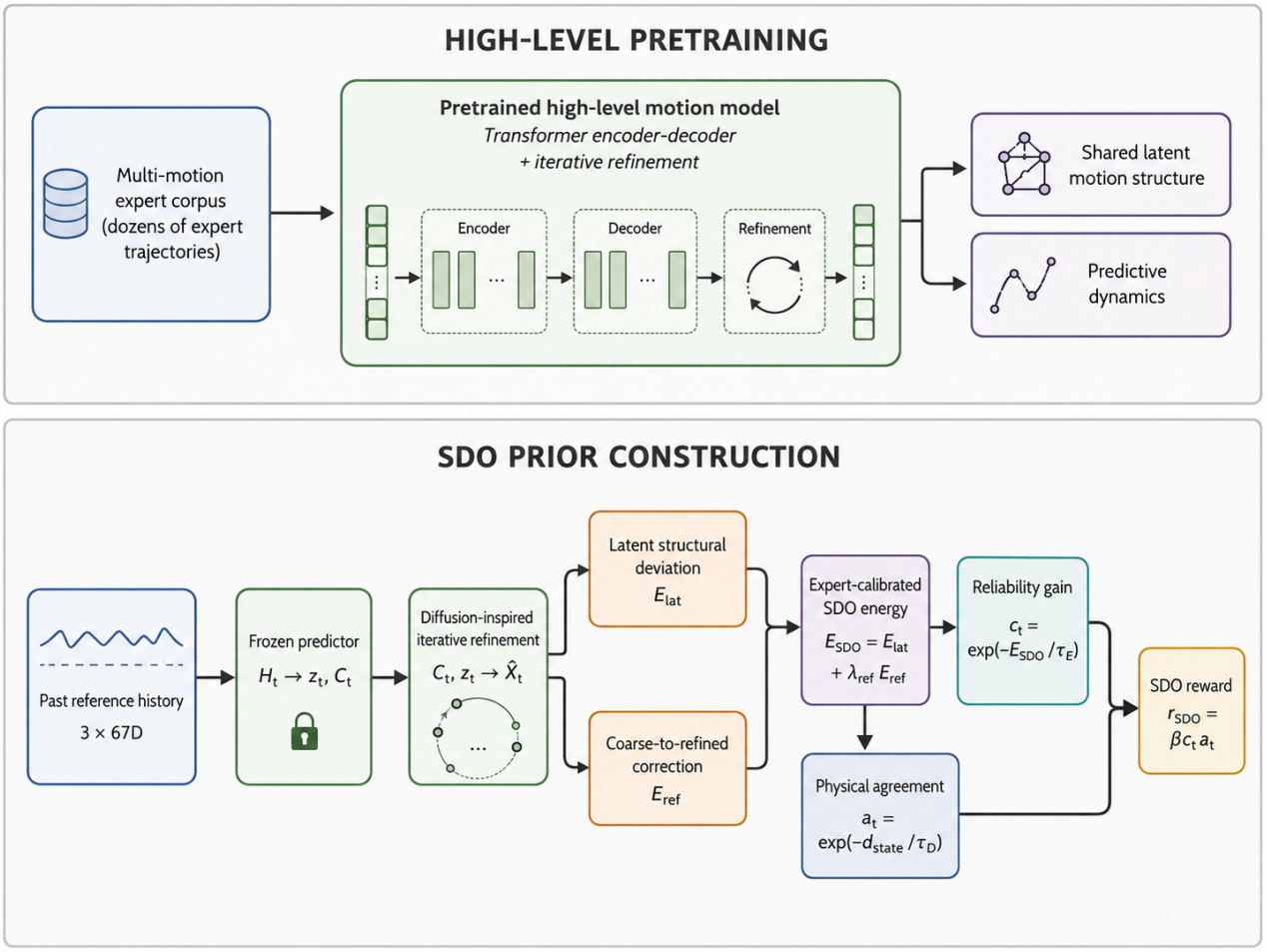

Fig. 1 SDO architecture. Multi-motion pretraining learns a shared Transformer representation and iterative predictive dynamics. After freezing, latent deviation and coarse-to-refined disagreement are calibrated into a reliability gain; together with physical agreement, this gain modulates a weak SDO reward added to the original imitation objective

3.2 Baseline reinforcement-learning problem

We use a trajectory-conditioned Markov decision process

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma), \tag{5}$$

where s_t is the simulator state and a_t is the torque-control action. The policy $\pi_\psi(a_t \mid o_t)$ is optimized by standard PPO [22]. Its policy architecture, rollout path, clipped surrogate objective, generalized advantage estimation, action interface, observation interface, and optimizer are not modified.

Let $r_{\text{mimic}}(t)$ be the original LocoMuJoCo MimicReward. SDO changes only the training reward:

$$r_{\text{total}}(t) = r_{\text{mimic}}(t) + r_{\text{SDO}}(t). \tag{6}$$

With $\beta = 0$, the SDO term is exactly zero and Eq. (6) reduces to the baseline. This property is also checked in the MJX integration tests.

3.3 Multi-motion Transformer representation and coarse future

The high-level predictor is trained on a pooled expert dataset rather than a single formal control trajectory. Every valid trajectory contributes normalized history/future windows to the same training set, allowing the model to reuse regularities that recur across locomotion and posture changes. This is the source of the "structured" part of SDO: z_t is not a code learned only for the downstream PPO reference, but a coordinate produced by a model fitted across a wider expert-motion repertoire.

Each of the three 67-dimensional history frames is projected to a 512-dimensional model space and receives sinusoidal positional encoding. Six Pre-LayerNorm Transformer encoder layers process the history; six decoder

layers with eight attention heads and a 2048-dimensional feed-forward block decode a learned bank of thirty future queries. The encoder memory is flattened and passed through LayerNorm, a 1024-unit GELU layer, and a 256-dimensional projection to form z_t. The decoded future passes through a residual MLP and a 512-to-67 projection to form C_t.

The coarse predictor is trained with

$$\mathcal{L}_{\text{coarse}} = \frac{1}{MD}\|C_t - Y_t\|_F^2, \qquad M = 30, \quad D = 67, \tag{7}$$

where Y_t is the normalized 30-frame expert future paired with history H_t. The future sequence is a supervised target during high-level pretraining; it is not an input when a trajectory-bound SDO prior is later constructed.

3.4 Diffusion-inspired dynamic refinement

After coarse-predictor training, the Transformer is frozen. The second stage receives (C_t, z_t) and applies twenty reverse-time, time-conditioned correction steps. A cosine schedule supplies step coefficients; each refinement step concatenates the current future estimate, the 256-dimensional latent code, and a 256-dimensional time embedding. Six residual MLP blocks process this concatenated representation and project it back to the 67-dimensional motion state.

The implementation uses a v-like internal correction variable and the iterative structure of a diffusion sampler, but it is trained directly against the target future:

$$\hat{X}_t = g_\phi(C_t, z_t), \qquad \mathcal{L}_{\text{refine}} = \frac{1}{MD}\|\hat{X}_t - Y_t\|_F^2. \tag{8}$$

We therefore refer to this module as *diffusion-inspired iterative refinement*: it borrows a time-conditioned reverse-update structure, while its training objective remains direct future-sequence reconstruction. For formal SDO prior construction, stochastic perturbation inside the iterative loop is disabled so that the same history produces the same refined future and the same refinement energy.

The refiner has two jobs in the present method. It produces the final predictive target, and it exposes a dynamic discrepancy: the amount by which the complete future sequence moved away from the Transformer's initial forecast. That second quantity is the basis of E_{ref} below.

Table 1 High-level predictor used to construct the SDO prior

Component	Configuration
Motion state	34 qpos + 33 qvel = 67D
History length N	3 frames
Future horizon M	30 frames
Transformer dimension	512
Attention heads	8
Encoder / decoder layers	6 / 6
Feed-forward dimension	2048
Latent dimension	256
Iterative refinement steps	20
Formal prior inference	deterministic

3.5 Reference-history-conditioned offline prior

Once high-level pretraining is complete, the predictor is frozen. For the low-level task, SDO constructs a trajectory-bound prior from the reference history alone. At each valid timestep, the preceding three reference states are passed through the frozen model to obtain $(z_t, C_t, \hat{X}_t)$. No future ground-truth state is provided to the predictor at this stage. The first refined target, the two energy components, validity information, and fixed state scales are stored and indexed by the reference timestep.

PPO therefore does not run the Transformer or the iterative refiner inside its update graph. During rollout, the reward path retrieves the precomputed entry and compares the current simulated state with the predicted

target. Future expert states reappear only in two places where they legitimately belong: as supervised labels when the high-level predictor is pretrained and as held-out targets when prediction error is measured. They are not exposed as an additional future observation to the policy.

This separation is useful beyond efficiency. The multi-motion model carries reusable temporal knowledge into the downstream task, while the low-level learner remains an ordinary trajectory-conditioned controller. SDO is the interface between them: predictor-derived structure is converted to a scalar gain without requiring shared gradients, shared network layers, or a PPO-specific modification.

3.6 Expert calibration and latent structural energy

Reliability is measured against fixed expert statistics rather than statistics refitted on the formal target. Calibration uses designated expert trajectories to estimate the state mean and standard deviation, per-dimension latent mean μ_z and standard deviation σ_z, and robust quantiles of the raw energy distributions. The calibration artifact contains 20,834 valid history/future windows and remains frozen during validation and formal PPO training.

The raw latent deviation is

$$e_{\text{lat}}^{\text{raw}}(t) = \frac{1}{d_z} \sum_{j=1}^{d_z} \left(\frac{z_{t,j} - \mu_{z,j}}{\sigma_{z,j} + \epsilon} \right)^2, \qquad d_z = 256. \tag{9}$$

Equation (9) is used operationally as an expert-calibrated structural-deviation score: histories mapped farther from the concentrated expert representation receive larger latent energy. This role requires a stable ranking of structural atypicality rather than a probabilistic density model.

Both raw energies use the same robust normalization operator

$$\mathcal{R}(e; q_{50}, q_{95}) = \text{clip}\left(\frac{e - q_{50}}{q_{95} - q_{50} + \epsilon}, 0, 3 \right). \tag{10}$$

Thus

$$E_{\text{lat}}(t) = \mathcal{R}\left(e_{\text{lat}}^{\text{raw}}(t); q_{50}^{\text{lat}}, q_{95}^{\text{lat}} \right). \tag{11}$$

Median-to-95th-percentile scaling gives the gate a stable reference range without allowing a few calibration outliers to set the denominator.

3.7 Dynamic correction energy

The second signal is the predictor's inter-stage correction magnitude. In normalized state space,

$$e_{\text{ref}}^{\text{raw}}(t) = \frac{1}{MD} \| \hat{X}_t^{(n)} - C_t \|_F^2, \tag{12}$$

with $M = 30$ and $D = 67$. The corresponding calibrated value is

$$E_{\text{ref}}(t) = \mathcal{R}\left(e_{\text{ref}}^{\text{raw}}(t); q_{50}^{\text{ref}}, q_{95}^{\text{ref}} \right). \tag{13}$$

Equation (12) is interpreted first as *inter-stage disagreement*. The Transformer proposes C_t, whereas the learned refinement dynamics produce $\hat{X}_t$; a large displacement means that the future accepted by the second stage requires a substantial revision of the coarse temporal forecast. Correction magnitude alone does not establish unreliability, so its role is validated rather than assumed. On the validation data, larger E_{ref} tracks larger future prediction error, with the strongest reported association at the 15- and 30-step horizons. This empirical link is what makes refinement correction useful to SDO as a predictor-difficulty signal.

E_{lat} and E_{ref} are complementary. The first asks whether the conditioning history occupies a familiar region of the expert-calibrated representation; the second asks how strongly the predictor's refinement dynamics disagree with its own coarse future. A latent-familiar history can still require a large future correction, while an atypical history can occasionally yield coarse and refined futures that agree. Combining the two prevents either form of internal consistency from being mistaken for reliability on its own.

3.8 From energy to reliability gain

The aggregation in Eq. (4) follows a compositional requirement rather than an arbitrary stacking of scores. After robust calibration, both E_{lat} and E_{ref} are non-negative evidence against trusting the proposed future. We require zero energy to preserve unit gain and independent evidence to compound attenuation. For a positive continuous gain map g, the relation $g(E_1 + E_2) = g(E_1)g(E_2)$ with $g(0) = 1$ gives the exponential family. Applying it to the two calibrated signals yields

$$c_t = \exp\left(-\frac{E_{\text{lat}}(t)}{\tau_{\text{energy}}}\right)\exp\left(-\frac{\lambda_{\text{ref}}E_{\text{ref}}(t)}{\tau_{\text{energy}}}\right) = \exp\left(-\frac{E_{\text{SDO}}(t)}{\tau_{\text{energy}}}\right), \quad 0 < c_t \le 1. \tag{14}$$

This gives the additive energy in Eq. (4) and the exponential gain in one construction: each discrepancy can reduce trust, while neither can increase the gain above one. Because both energies already share the same robust calibration scale, λ_{ref} controls relative sensitivity rather than unit conversion. The formal value $\lambda_{\text{ref}} = 2.0$ and the temperature are selected on the validation subset and locked before the formal study.

Equation (14) is used as a *reliability proxy*: the needed empirical property is that larger calibrated energy tends to accompany larger future prediction error. Section 5.1 tests that ordering directly.

3.9 Policy–prior agreement: the direction of modulation

Energy alone says whether a prediction deserves trust; it does not say whether the current simulated state is near the predicted motion. To make the two states unambiguous, write the simulator state as $x_t^{\text{sim}} = [q_t^{\text{sim}}, \dot{q}_t^{\text{sim}}]$ and the refined prior target as $\hat{x}_t = [\hat{q}_t, \hat{\dot{q}}_t]$. The target then supplies the local physical direction through a structured state-distance kernel.

For non-quaternion generalized positions,

$$d_{\text{pos}}(t) = \frac{1}{|\mathcal{I}_p|}\sum_{i\in\mathcal{I}_p}\left(\frac{q_{t,i}^{\text{sim}} - \hat{q}_{t,i}}{s_{q,i} + \epsilon}\right)^2, \tag{15}$$

and for generalized velocities,

$$d_{\text{vel}}(t) = \frac{1}{n_v}\sum_{i}\left(\frac{\dot{q}_{t,i}^{\text{sim}} - \hat{\dot{q}}_{t,i}}{s_{\dot{q},i} + \epsilon}\right)^2. \tag{16}$$

Quaternion orientation is handled separately with the sign-invariant angular metric. For the set $\mathcal{Q}$ of quaternion blocks,

$$d_{\text{rot}}(t) = \frac{1}{|\mathcal{Q}|}\sum_{j\in\mathcal{Q}} 2\arccos\left(\left|\left(u_{t,j}^{\text{sim}}\right)^{\top}\hat{u}_{t,j}\right|\right), \tag{17}$$

where each quaternion is normalized before the inner product. Position, velocity, and rotation are treated as separable agreement channels. If each channel contributes a factor $\exp(-w_k d_k/\tau_{\text{distance}})$, multiplying the three factors gives an additive distance in the exponent,

$$d_{\text{state}} = w_{\text{pos}}d_{\text{pos}} + w_{\text{vel}}d_{\text{vel}} + w_{\text{rot}}d_{\text{rot}}, \tag{18}$$

and therefore

$$a_t = \exp\left(-\frac{d_{\text{state}}(t)}{\tau_{\text{distance}}}\right), \qquad 0 < a_t \le 1. \tag{19}$$

The formal configuration uses $w_{\text{pos}} = w_{\text{vel}} = w_{\text{rot}} = 1$. Since position and velocity errors are already normalized by fixed expert-derived scales and rotation uses its own normalized angular metric, equal weights provide a neutral channel balance; τ_{distance} controls the overall sharpness of the agreement gate.

This separation is central to SDO. The predictor target determines the direction in physical state space; c_t controls whether that direction deserves influence; a_t controls whether the policy is currently compatible with it. No single one of these quantities is asked to perform all three roles.

3.10 Weak optimization modulation

The final SDO reward is

$$r_{\mathrm{SDO}}(t) = \beta c_t a_t = \beta \exp\left(-\frac{E_{\mathrm{SDO}}(t)}{\tau_{\mathrm{energy}}}\right) \exp\left(-\frac{d_{\mathrm{state}}(t)}{\tau_{\mathrm{distance}}}\right). \tag{20}$$

The formal setting uses $\beta = 0.05$, $\tau_{\mathrm{energy}} = 1.0$, and $\tau_{\mathrm{distance}} = 0.5$. Invalid prior entries are masked to zero. Since both gates are at most one,

$$0 \le r_{\mathrm{SDO}}(t) \le \beta, \tag{21}$$

so the auxiliary signal is explicitly bounded and cannot numerically replace the primary imitation objective.

If reliability gating is removed by setting $c_t = 1$, Eq. (20) becomes

$$r_{\mathrm{ungated}}(t) = \beta a_t, \tag{22}$$

which is an ordinary predictive tracking bonus. This identity makes clear where the contribution lies. The reward interface itself is simple by design; SDO's new information is the predictor-derived, expert-calibrated gain that distinguishes Eq. (20) from Eq. (22). Figure 2 illustrates the two gates and their product.

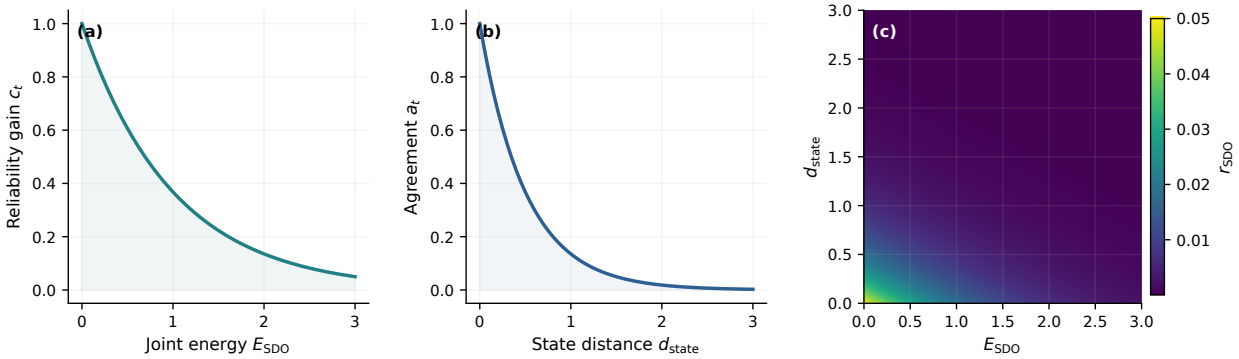

Fig. 2 The two SDO gates under the formal parameters. (a) Expert-calibrated joint energy is mapped monotonically to reliability gain. (b) State distance is mapped to policy–prior agreement. (c) Their product yields a bounded auxiliary reward; strong guidance is concentrated in the low-energy, low-distance region

Algorithm 1: Offline construction and runtime use of SDO. For each reference timestep, take only the preceding three reference states and run the frozen high-level predictor to obtain $(z_t, C_t, \hat{X}_t)$. Compute raw latent deviation and full-horizon coarse-to-refined correction, normalize both using fixed expert calibration, and store $\hat{x}_t$, E_{lat}, E_{ref}, and validity metadata in the trajectory-bound prior. During PPO rollout, retrieve those entries, compute current-state agreement a_t, form $r_{\mathrm{SDO}} = \beta c_t a_t$, and add it to MimicReward. No high-level network inference or SDO gradient is inserted into the PPO loss.

3.11 What SDO adds beyond a predictive tracking reward

The final reward in Eq. (20) is intentionally simple, so the distinction between SDO and ordinary reward shaping is best stated in terms of information flow. A frame-wise imitation reward has access to the current reference. An ungated predictive bonus additionally has access to a predicted target. SDO adds a third kind of information: evidence produced inside a multi-motion predictor about how that target was constructed. The latent term and the refinement term are therefore not extra tracking errors. They are properties of the frozen high-level model before the low-level policy is compared with the target.

Table 2 Information used by the three reward constructions considered in this study

Information source	Frame-wise imitation	Ungated predictive bonus	SDO
Current reference state	Used	Used through the base reward	Used through the base reward
Predicted future target	–	Used directly	Used as the motion direction
Latent structural deviation	–	–	Controls target gain
Coarse-to-refined correction	–	–	Controls target gain
Current policy–prior agreement	–	Used	Used separately from reliability

This separation matters because prediction error and control error need not be the same event. A policy may be close to a target that was generated from an atypical history, while a low-energy target may initially be far from the policy because the physical controller has not yet reached it. Collapsing both situations into one state distance makes the prior's own reliability invisible. SDO keeps the two questions separate: the high-level model determines how much authority the direction deserves, and the low-level state determines how well the controller currently agrees with that direction. The resulting mechanism remains lightweight at runtime, but the gain carries information learned from the wider expert-motion corpus rather than from the single downstream control reference alone.

4 Experimental design

4.1 Environment and control task

Experiments use a standard physics-based humanoid imitation setup [13,23] with established expert motion references [24]. The low-level controller retains the original trajectory-based imitation reward. The formal control trajectory contains coupled turning and walking, making it a useful test of whether a weak high-level bias can remain compatible with balance-sensitive torque control. SDO adds neither policy observations nor actions. Prior loading verifies consistency of the state representation and orientation handling before training starts.

4.2 Four data roles

The experiment separates four roles that are easy to conflate (Table 3). The high-level predictor is pretrained on a broad multi-motion corpus containing dozens of expert trajectories. A designated expert calibration subset then fixes latent statistics, state scales, and energy quantiles; it does not retrain the predictor. A separate validation subset is used for SDO reward-parameter selection and energy–prediction analysis. Finally, the paired 300M-step PPO study is performed on the formal turning-and-walking control trajectory. Internal dataset filenames are implementation details and are not part of the method definition.

Table 3 Data roles in the SDO pipeline

Stage	Role
High-level pretraining	Learn shared 67D temporal structure and predictive dynamics from dozens of expert trajectories
Expert calibration	Fix state scales, latent statistics, and robust energy quantiles; 20,834 valid calibration windows
Validation	Select SDO reward parameters and analyze energy–prediction association
Formal control study	Paired five-seed baseline versus SDO training on a challenging turning-and-walking reference

4.3 Validation and locked parameters

Validation uses deterministic return under the same evaluator later used for the formal comparison, with $\beta_{\text{eval}} = 0$. A coarse-to-fine sweep selects

$$\beta = 0.05, \qquad \lambda_{\text{ref}} = 2.0, \qquad \tau_{\text{energy}} = 1.0, \qquad \tau_{\text{distance}} = 0.5. \tag{23}$$

The selected setting reaches a three-seed validation return of 19.2542 ± 0.4948. Figure 3 shows the repeated three-seed candidates used to choose the final configuration. These values are locked before the 300M-step runs.

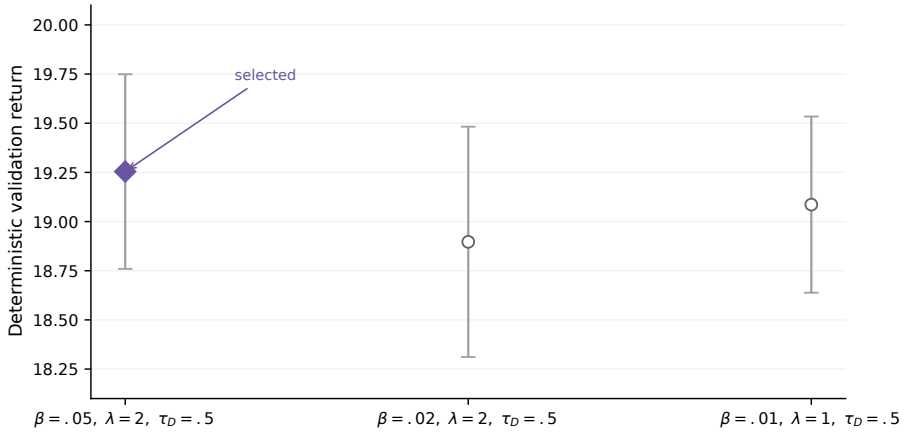

Fig. 3 Repeated three-seed candidates used during reward-parameter selection. The diamond marks the configuration locked before the formal control study

4.4 Formal PPO configuration and common evaluator

Each baseline/SDO pair uses the same random seed from $\{0, 1, 2, 3, 4\}$ and the same 300-million-step budget. PPO runs 2048 parallel environments with 200 rollout steps. The policy MLP has hidden layers of 512 and 256 units with tanh activations. The learning rate is 10^{-4}, $\gamma = 0.99$, GAE $\lambda = 0.95$, PPO clipping is 0.2, and each update uses four epochs and 32 minibatches. The low-level algorithm is identical in the two conditions.

Table 4 Formal PPO and SDO settings

PPO parameter	Value	SDO parameter	Value
Total timesteps	300M	β	0.05
Parallel environments	2048	λ_{ref}	2.0
Rollout steps	200	τ_{energy}	1.0
Learning rate	1×10^{-4}	τ_{distance}	0.5
γ	0.99	$w_{\text{pos}}, w_{\text{vel}}, w_{\text{rot}}$	1,1,1
GAE λ	0.95	Energy clip	3.0
PPO clip	0.20	Prior inference	deterministic
Update epochs	4		
Minibatches	32		

Training return is not used for the formal method comparison because SDO adds a positive training-time term. Saved policies are instead evaluated with deterministic actions, 32 parallel environments, 1000 steps, frozen running statistics, and $\beta_{\text{eval}} = 0$. Both methods are therefore scored only by the original MimicReward. The training seed, rather than the parallel evaluation environments, is the independent statistical unit.

Server validation was performed on an NVIDIA RTX 5090 with JAX 0.6.2 and MuJoCo 3.2.7. Pre-formal checks included 37 passing unit tests, two passing MJX integration tests, finite observation/action/reward checks, prior-schema validation, and direct $\beta = 0$ equivalence of the reward path.

4.5 Energy–prediction analysis

Prediction quality is evaluated at 1, 5, 15, and 30 steps on the validation subset. Composite error combines normalized non-quaternion qpos error, normalized qvel error, and quaternion angular error. Spearman correlation is used because the gate needs a useful ordering of difficult predictions, not a linear probability calibration.

4.6 Ungated-prior diagnostic

The code also supports the same predictive target with confidence fixed to one, corresponding to Eq. (22). A one-seed, 1M-step diagnostic compares baseline PPO, this ungated target, and full SDO. It is deliberately used for mechanism attribution rather than as a substitute for the five-seed formal comparison: the question is whether the same target behaves differently once predictor-derived energy is allowed to control its gain.

5 Results

5.1 Energy ranks prediction difficulty

The energy signal behaves as SDO requires it to behave. Joint SDO energy has Spearman correlation of approximately 0.29, 0.34, 0.38, and 0.34 with composite future prediction error at 1, 5, 15, and 30 steps (Fig. 4). The association strengthens toward the medium horizon and remains positive at 30 steps. This is enough for the reliability map in Eq. (14): predictions with larger internal discrepancy tend to be more difficult, so their influence can be reduced without interpreting energy as a calibrated probability.

The coarse-to-refined term carries the clearest long-horizon signal. E_{ref} reaches a Spearman correlation of about 0.55 with composite error at both 15 and 30 steps. A large full-horizon correction is therefore not just an internal numerical event; it is informative about the difficulty of the future the predictor is proposing. The validation-selected weight $\lambda_{\text{ref}} = 2$ gives this dynamic discrepancy meaningful leverage in the joint energy.

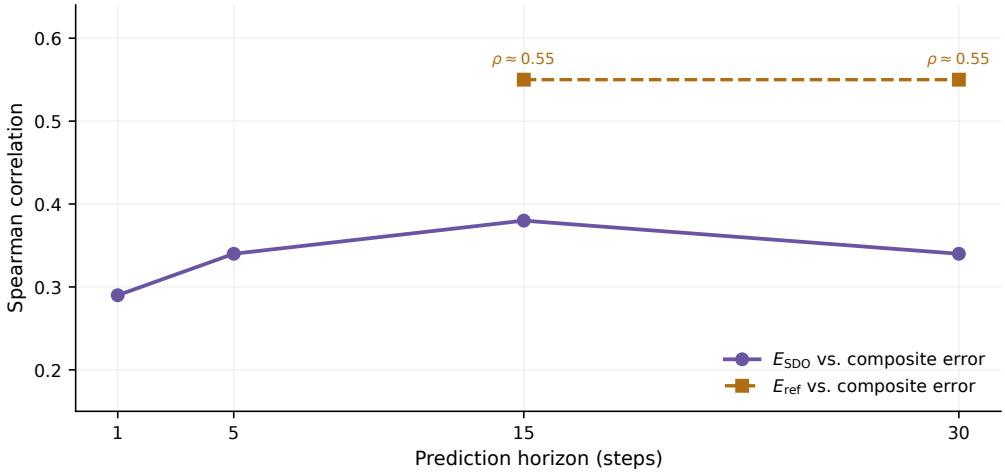

Fig. 4 Spearman association between predictor-derived energy and composite future-state error. Joint SDO energy remains positively associated with error from 1 to 30 steps; refinement energy reaches about 0.55 at the 15- and 30-step horizons

5.2 Why the energy gate matters

The short diagnostic isolates the difference between possessing a predictive target and using that target selectively. After approximately one million training steps, final training returns are 9.2903 for baseline, 9.1312 for the ungated predictive prior, and 9.3753 for full SDO. Under the common deterministic evaluator with $\beta_{\text{eval}} = 0$, the corresponding returns are approximately 20.365, 19.379, and 20.346 (Fig. 5).

The useful contrast is between the two predictive conditions. They share the target and low-level learner; what changes is whether predictor-internal energy controls the target's gain. Removing that gate produces a sizeable drop in the common evaluator, while restoring SDO brings performance back from 19.379 to 20.346, close to the 20.365 baseline. This diagnostic explains why the contribution is not the presence of an extra future target by itself.

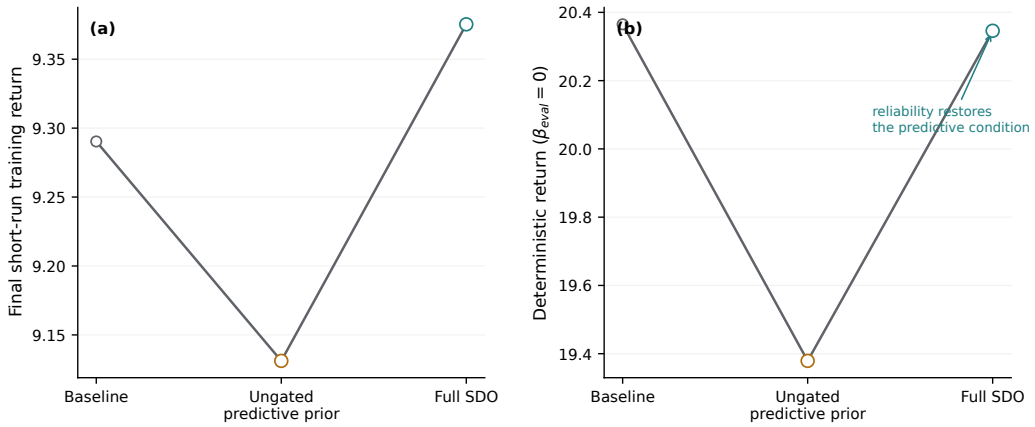

Fig. 5 Short diagnostic with a common SDO-free evaluator. The ungated predictive target lowers deterministic return, while predictor-derived reliability restores the result close to the baseline using the same target

5.3 Five paired 300M-step control runs

The formal comparison shows a uniform paired improvement. Baseline PPO reaches 16.4396 ± 0.0649 deterministic return, while SDO-guided training reaches 16.9191 ± 0.3574. The mean paired increase is +0.4795, or about +2.92%, and all five independently seeded SDO runs finish above the baseline trained with the same seed (Table 5). Episode duration moves in the same direction: the mean rises from 125.75 ± 2.20 to 134.61 ± 1.30, again with positive paired differences in all five seeds.

Table 5 Paired deterministic evaluation with $\beta_{\text{eval}} = 0$

Seed	Baseline	SDO	Difference	Relative difference
0	16.3528	16.8431	+0.4903	+2.998%
1	16.5289	17.2013	+0.6724	+4.068%
2	16.4688	16.6404	+0.1717	+1.042%
3	16.4242	17.3709	+0.9466	+5.764%
4	16.4235	16.5400	+0.1165	+0.709%
Mean	16.4396	16.9191	+0.4795	+2.917%

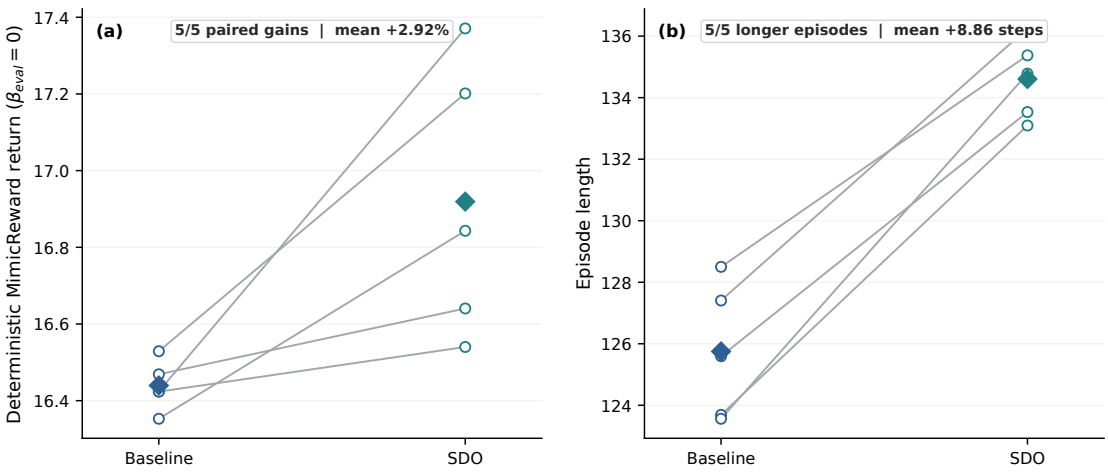

Fig. 6 Paired formal results for five independent training seeds. (a) Deterministic MimicReward return with the SDO bonus disabled for both methods. (b) Episode length. Every paired SDO run is higher on both measures; diamonds show method means

The paired return difference has standard deviation 0.3472 and standardized effect size $d_z = 1.381$. The

paired t test gives $p = 0.0366$ with a 95% confidence interval of $[0.048, 0.911]$ for the mean paired difference. The two-sided exact Wilcoxon signed-rank test gives $p = 0.0625$; with five nonzero pairs all moving in the same direction, this is the minimum attainable two-sided exact value at $n = 5$. Together with the 5/5 positive paired differences, these statistics support the same measured conclusion: the SDO-trained policies retain a modest but consistent advantage under the common evaluator.

That evaluator removes the most obvious alternative explanation. During evaluation, SDO-trained policies receive no reward for latent energy, refinement energy, confidence, or predictive-target agreement. Their scores are computed by the original MimicReward, exactly as for baseline PPO. The improvement therefore reflects behavior retained by the trained policy rather than an additional test-time reward term.

6 Discussion

SDO functions as gain control for a reusable motion prior. The high-level model has already learned from many expert trajectories before PPO starts. That matters because the downstream prior is not merely a smoothed copy of one control reference; it is produced by a model whose latent and predictive structure were shaped by a wider motion repertoire. SDO exploits information that such a model already computes but that ordinary imitation rewards discard.

Latent deviation and refinement correction are complementary for this purpose. E_{lat} is tied to the conditioning history and asks whether the recent motion sits in a familiar part of the learned representation. E_{ref} is tied to future construction and asks how far the predictor had to move away from its own coarse forecast. The latter becomes especially informative at longer horizons, where the 0.55 correlation with future error suggests that the refiner's correction carries a practical signature of prediction difficulty. Combining the two in energy space lets either kind of mismatch reduce the authority of the target.

This also clarifies the relation between the original energy–manifold motivation and the implemented algorithm. A literal formulation could attempt to push the policy along $-\nabla E$ in latent space. The present realization avoids that coupling. The refined target supplies the physical direction, and the energy supplies only its gain. The separation is useful because the low-level learner remains responsible for satisfying dynamics, contacts, and the original imitation objective; the high-level prior only biases that learning when its own internal evidence is favorable.

The ungated diagnostic isolates the need for selectivity: the same predictive target lowers the common evaluator when it is weighted uniformly, while predictor-derived reliability restores the result. The formal runs extend that observation under a much larger training budget: mean deterministic return rises by about 2.9%, with every paired seed moving upward, and episode duration increases in every pair as well. The magnitude is consistent with the intended operating regime: SDO is bounded by β and biases an already dense MimicReward rather than replacing it. The key observation is the direction and consistency of the learned-policy change under a common evaluator, not a change in the test-time objective.

The closest motion-prior families emphasize different interfaces. Adversarial priors such as AMP score motion plausibility; ASE and PULSE expose reusable skill spaces; FLD structures latent dynamics and can fall back from risky targets; SMP turns a frozen generative model into a reusable reward; and CMP adapts prior relevance to task context. SDO does not compete with those capabilities directly. Its contribution is the narrower predictor–controller interface: a generated target carries an explicit, prediction-specific gain derived from how the frozen predictor reached that target. Because this gain enters only through the reward layer, PPO provides the concrete test bed without defining the method; another low-level optimizer could consume the same state, target, and reliability signal without inheriting a PPO-specific update term.

7 Conclusion

Structured Dynamics Optimization carries knowledge from a multi-motion predictive model into physics-based reinforcement learning without turning the predictor into the controller. A Transformer/refinement model is first pretrained on dozens of expert trajectories. Its latent deviation and coarse-to-refined correction are then calibrated into a joint energy, and that energy decides how strongly a predicted motion target is allowed to influence the low-level learner. The target gives direction; energy gives gain; policy–prior agreement supplies the final physical check.

Mechanism and control evidence align along the same chain. Predictor-derived energy ranks future prediction difficulty; the ungated diagnostic isolates the need for selective guidance; and the full five-seed study retains higher return and longer episodes under a common evaluator after the auxiliary term has been removed. SDO raises deterministic imitation return by about 2.9% on average, with a positive paired difference in every seed. This is the behavior expected from a weak modulation term: the base imitation objective remains dominant, while predictor-derived structure contributes a consistent selective bias. The contribution is the interface itself–extracting structural and dynamic discrepancy from a pretrained motion prior and converting it into a selective optimization signal for an otherwise unchanged reinforcement-learning controller.

Statements and Declarations

Funding. The author received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors for this work.

Competing interests. The author declares no financial or non-financial competing interests that are directly or indirectly related to this work.

Author contributions. Ange Tong: conceptualization, methodology, software, validation, formal analysis, visualization, investigation, writing - original draft, and writing - review and editing.

Data availability. Source code and public configuration are available at https://github.com/Alfred-Duncan/sdo. The repository provides the predictive-prior, calibration, energy, and reward-integration code; large datasets, trained checkpoints, generated priors, training logs, and experiment artifacts are not distributed.

Ethics approval. Not applicable. The reported study uses physics simulation and existing motion datasets and does not involve new experiments with human participants or animals.

References

[1] Peng XB, Abbeel P, Levine S, van de Panne M (2018) DeepMimic: example-guided deep reinforcement learning of physics-based character skills. ACM Trans Graph 37(4):143. https://doi.org/10.1145/3197517.3201311
[2] Peng XB, Ma Z, Abbeel P, Levine S, Kanazawa A (2021) AMP: adversarial motion priors for stylized physics-based character control. ACM Trans Graph 40(4).
[3] Peng XB, Guo Y, Halper L, Levine S, Fidler S (2022) ASE: large-scale reusable adversarial skill embeddings for physically simulated characters. ACM Trans Graph 41(4).
[4] Luo Z, Cao J, Winkler A, Kitani K, Xu W (2023) Perpetual humanoid control for real-time simulated avatars. In: Proc IEEE/CVF Int Conf Comput Vis, pp 10895–10904.
[5] Luo Z, Cao J, Merel J, Winkler A, Huang J, Kitani K, Xu W (2024) Universal humanoid motion representations for physics-based control. In: International Conference on Learning Representations.
[6] Tessler C, Guo Y, Nabati O, Chechik G, Peng XB (2024) MaskedMimic: unified physics-based character control through masked motion inpainting. ACM Trans Graph 43(6). https://doi.org/10.1145/3687951
[7] Li C, Stanger-Jones E, Heim S, Kim S (2024) FLD: Fourier latent dynamics for structured motion representation and learning. In: International Conference on Learning Representations.
[8] Jiang Z, Xu Y, Wagener N, Luo Y, Janner M, Grefenstette E, Rocktäschel T, Tian Y (2024) H-GAP: humanoid control with a generalist planner. In: International Conference on Learning Representations.
[9] Zhao W, Zhao Y, Pajarinen J, Muehlebach M (2025) Bi-Level Motion Imitation for Humanoid Robots. Proc Mach Learn Res 270:963–981.
[10] Park S, Lee E, Lee KB, Lee SH (2026) MaskAdapt: learning flexible motion adaptation via mask-invariant prior for physics-based characters. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit, pp 2285–2294.
[11] Mu Y, Zhang Z, Shi Y, Matsumoto M, Imamura K, Tevet G, Guo C, Taylor M, Shu C, Xi P, Peng XB (2025) SMP: Reusable Score-Matching Motion Priors for Physics-Based Character Control. arXiv:2512.03028.
[12] Mo Y, Gu Y, Zhou Y, Cao H, Xu R (2026) Learning Context-Aware Motion Priors for Humanoid Control. arXiv:2608.03234.
[13] Al-Hafez F, Zhao G, Peters J, Tateo D (2023) LocoMuJoCo: a comprehensive imitation learning benchmark for locomotion. arXiv:2311.02496.
[14] Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30.
[15] Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851.
[16] Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH (2023) Human Motion Diffusion Model. In: International Conference on Learning Representations.
[17] Yuan Y, Song J, Iqbal U, Vahdat A, Kautz J (2023) PhysDiff: physics-guided human motion diffusion model. In: Proc IEEE/CVF Int Conf Comput Vis, pp 16010–16021. https://doi.org/10.1109/ICCV51070.2023.01467
[18] Kendall A, Gal Y (2017) What uncertainties do we need in Bayesian deep learning for computer vision? Adv Neural Inf Process Syst 30.
[19] Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst 30.
[20] Lee K, Lee K, Lee H, Shin J (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Adv Neural Inf Process Syst 31.
[21] Liu W, Wang X, Owens J, Li Y (2020) Energy-based out-of-distribution detection. Adv Neural Inf Process Syst 33:21464–21475.
[22] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv:1707.06347.
[23] Todorov E, Erez T, Tassa Y (2012) MuJoCo: a physics engine for model-based control. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp 5026–5033. https://doi.org/10.1109/IROS.2012.6386109
[24] Harvey FG, Yurick M, Nowrouzezahrai D, Pal C (2020) Robust motion in-betweening. ACM Trans Graph 39(4). https://doi.org/10.1145/3386569.3392480