ECHO: IN-CONTEXT BEHAVIOR FOUNDATION MODEL FOR HUMANOID WHOLE-BODY CONTROL

Anonymous authors
Paper under double-blind review

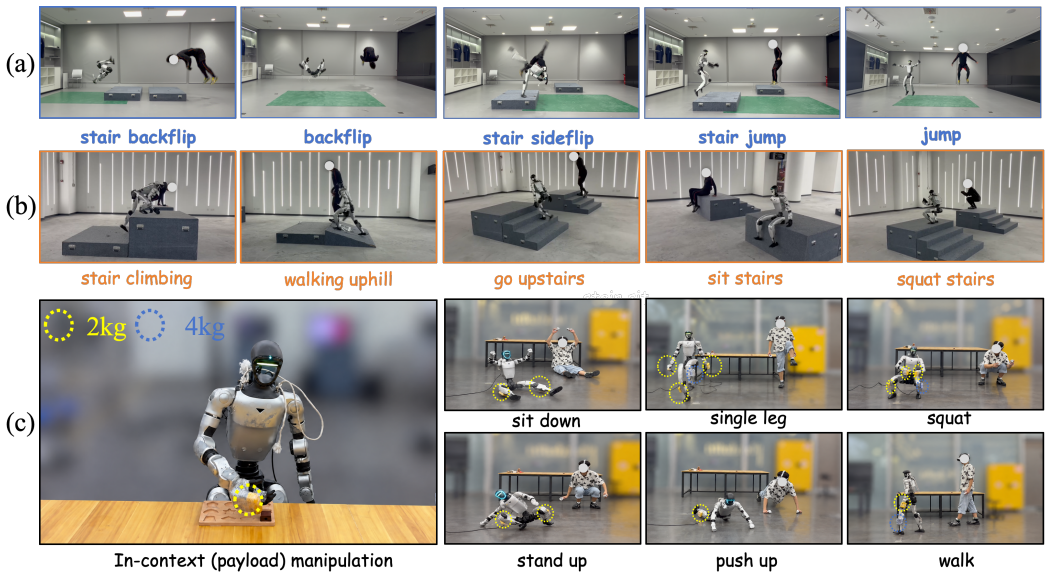

Figure 1: **ECHO: Broad whole-body capabilities with in-context adaptation.** A shared behavior prior supports whole-body teleoperation spanning extreme acrobatics (a) and terrain-interactive behaviors (b), while interaction-derived context enables adaptation to payload-induced dynamics changes during manipulation and whole-body control (c).

ABSTRACT

Humanoid behavior foundation models (BFMs) acquire diverse skills through large-scale motion imitation and massively parallel reinforcement learning, yet broad behavioral capability does not ensure precise execution across physical conditions. Without distinguishing the underlying dynamics, policies may favor conservative compromises over environment-specific precision. We introduce **ECHO**, an in-context BFM that turns interaction experience into environment-specific control, separating general behavioral capability from dynamics-specific execution. Mixed and Annealed Sampling first establishes a behavior prior in a single pretraining stage by balancing targeted practice with broad motion coverage. Building on this prior, Context Representation and Residual Adaptation learns structured dynamics representations from interaction histories to guide residual corrections to the frozen prior, enabling adaptation without test-time parameter updates. Extensive experiments show that ECHO outperforms all baseline methods across diverse motions and physical conditions. Moreover, deployment on the Unitree G1 demonstrates whole-body teleoperation spanning extreme acrobatics and terrain interaction, alongside payload adaptation in manipulation and whole-body control. Page is available at https://echo-bfm.github.io/echoweb/.

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.

1 INTRODUCTION

For the things we have to learn before we can do them, we learn by doing them.

Aristotle, *Nicomachean Ethics*

Humanoid robots are expected to execute diverse whole-body behaviors accurately, responsively, and robustly in human-centric environments. Large-scale motion imitation and massively parallel reinforcement learning (RL) have advanced this goal through behavior foundation models (BFMs) (Luo et al., 2026; Zeng et al., 2025; Qi et al., 2026; Zeng et al., 2026; RoboParty Lab Team, 2026; Liu et al., 2026a; Cheng et al., 2026; Ma et al., 2026; Li et al., 2026; Chen et al., 2026b; Liu et al., 2026b). Yet a broad behavior repertoire alone does not ensure precise execution across physical conditions: the same reference motion may require different control responses as payload, friction, or actuator dynamics vary. Without adequately distinguishing these dynamics, a policy trained across environments may learn conservative compromises that favor broad reliability over environment-specific tracking precision (Mehta et al., 2020; Wang et al., 2026b). The key challenge is therefore to enable a BFM to understand the underlying dynamics across environments and adapt to different physical conditions.

Recent work in robotic manipulation has demonstrated in-context adaptation through human demonstrations (Shah et al., 2025) or retrieved interaction experience (Chen et al., 2026a; Zhang et al., 2025a), without updating the model weights at test time. These advances motivate equipping humanoid BFMs with the ability to learn environmental dynamics from interaction context and tailor whole-body actions accordingly. As a BFM executes diverse behaviors, its actions and the resulting state transitions provide evidence about how the underlying dynamics shape its response. By encoding this evidence into structured dynamics representations, the model can learn control-relevant properties of the current environment and use them to guide subsequent actions. Interaction histories thus serve not merely as records of past execution, but as context for environment-specific action corrections. This allows a shared behavior prior to pursue the same objective with control responses tailored to different physical conditions, adapting its execution without test-time parameter updates.

To this end, we introduce **ECHO**, an in-context behavior foundation model for humanoid whole-body control that separates general behavioral capability from dynamics-specific execution. Specifically, we first train a motion-tracking prior, **ECHO-PRE**, using Mixed and Annealed Sampling (MIAS). By balancing targeted practice with broad motion coverage, MIAS enables a single policy to jointly learn everyday motions, highly dynamic skills (Wang et al., 2026c), contact-rich floor behaviors, and terrain motions within one pretraining stage (Table 1). we introduce Context Representation and Residual Adaptation (CRA), which learns structured representations of the underlying environmental dynamics from the prior's interaction histories. CRA combines multi-step dynamics prediction with physical neighborhood soft alignment to capture control-relevant physical information across behaviors. A context-conditioned residual policy then uses these representations to generate action corrections tailored to the learned dynamics, adapting the frozen prior's execution.

We evaluate ECHO from four fundamental perspectives: (1) *tracking fidelity*, covering challenging high-dynamic skills (Figure 1 (a)) and terrain motion (Figure 1 (b)); (2) *in-context adaptation*, assessing the benefit of interaction context under nominal and heavily randomized dynamics; (3) *representation quality*, examining whether representations distinguish physical environments across motions; and (4) *real-world applicability*, demonstrating whole-body teleoperation and payload adaptation on the Unitree G1 (Figure 1 (c)). Extensive experimental results show that MIAS strengthens the general behavior prior, while CRA learns physically informative context that supports more precise execution across varying dynamics without test-time parameter updates.

Our contributions are summarized as follows:

- We introduce **ECHO**, an in-context behavior foundation model that separates general behavioral competence from environment-specific execution, enabling a single whole-body behavior prior to retain broad capabilities while adapting to varying physical dynamics.
- We develop two complementary components: **Mixed and Annealed Sampling (MIAS)** for acquiring a broad whole-body repertoire in a single pretraining stage, and **Context**

Table 1: **Comparison of whole-body control capabilities.** ECHO-PRE covers all four capability categories within a single pretrained policy. ✓ indicates a demonstrated capability, while × indicates one not reported in the corresponding work.

	Extreme play	Extreme teleop	Ground interaction	Terrain interaction
SONIC (Luo et al., 2026)	✓	×	✓	×
GigaBrain-WBC-0.5 (Cheng et al., 2026)	✓	×	×	✓
HEFT (Liu et al., 2026a)	✓	×	×	×
Extreme-RGMT (Ma et al., 2026)	✓	✓	×	×
ECHO-PRE	✓	✓	✓	✓

Representation and Residual Adaptation (CRA) for turning interaction histories into physically grounded context that guides dynamics-specific corrections to the frozen prior.

- Extensive experiments show that ECHO improves tracking under substantial dynamics variation, learns context representations that remain physically informative across motions, and transfers to real-world whole-body teleoperation and payload adaptation.

2 Related Work

Behavior Foundation Models. Behavior foundation models learn reusable whole-body control priors. BFM Zeng et al. (2025) supports diverse control interfaces through generative behavioral representations, while BFM-Zero Li et al. (2026) uses unsupervised RL to unify motion-, goal-, and reward-conditioned control in a shared latent space. SONIC Luo et al. (2026), Scaling BFM Zeng et al. (2026), Humanoid-GPT Qi et al. (2026), and HoloMotion-1 Chen et al. (2026b) explore complementary data, model, and training scaling strategies to improve tracking and generalization, while Mimiclite RoboParty Lab Team (2026) emphasizes efficient generalized motion tracking. Beyond tracking, related systems support reactive motion planning (Chen et al., 2026c), terrain-aware execution (Wang et al., 2026e), and environmental interaction (Cheng et al., 2026). HEFT Liu et al. (2026a) and Extreme-RGMT Ma et al. (2026) further extend capabilities to heavy-payload teleoperation and highly dynamic skills, respectively. Our ECHO complements these advances by combining broad skill acquisition via MIAS with dynamics-specific residual adaptation through CRA.

In-Context Learning and Dynamics Adaptation. Algorithm Distillation Laskin et al. (2022) and Prompt-DT Xu et al. (2022) demonstrate adaptation through learning histories and demonstration prompts, respectively, without test-time parameter updates. In robotics, earlier work combines universal policies with online system identification to adapt control to unknown dynamics (Yu et al., 2017). RMA Kumar et al. (2021) and dual-history policies Li et al. (2025) exploit proprioceptive histories for dynamics adaptation, while DWL Gu et al. (2024) uses denoising reconstruction for implicit system identification. LocoFormer Liu et al. (2025) leverages cross-trial Transformer memory to adapt locomotion to unseen morphologies and dynamics. For humanoid control, Any2Track Zhang et al. (2025b) combines world-model-based dynamics representations with layer-wise adapters in a frozen tracker, while HoRD Wang et al. (2026a) integrates history-conditioned RL with online distillation. FADA Xie et al. (2026) separates planning from inverse dynamics and adapts the latter using target-domain rollouts. In contrast, ECHO combines multi-step dynamics prediction with cross-motion physical neighborhood soft alignment to learn structured context representations, then trains an RL-based residual policy (Johannink et al., 2019; Wang et al., 2026d) conditioned on the learned in-context representation with both the encoder and behavior prior frozen.

3 Method

As illustrated in Figure 2, we first formulate In-Context BFM under unobserved environmental dynamics (Sec. 3.1). We then introduce MIAS to pretrain ECHO-PRE, balancing broad motion coverage with targeted practice on difficult segments (Sec. 3.2). Building on this frozen prior, CRA enables effective in-context adaptation by inferring environmental dynamics from interaction history and translating the resulting context into action corrections tailored to the current environment (Sec. 3.3).

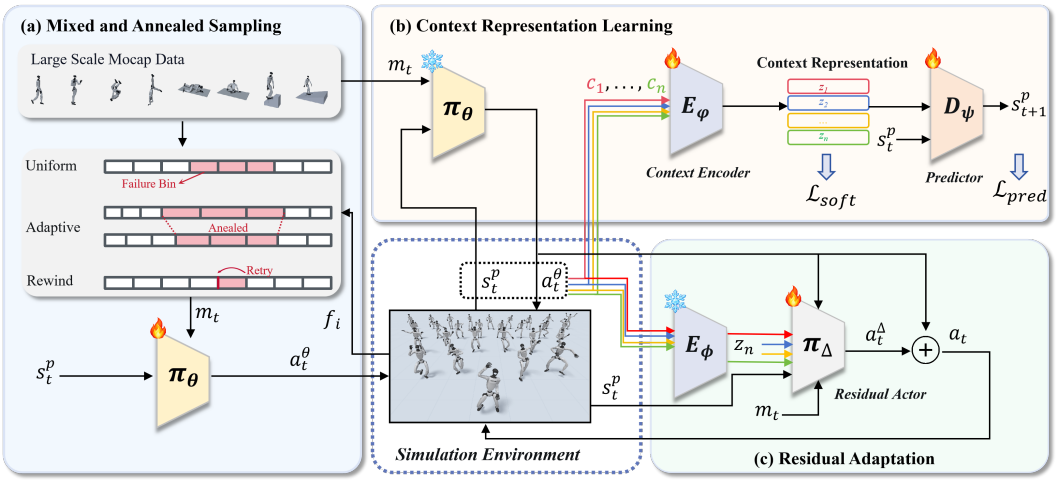

Figure 2: **Method Overview. ECHO-PRE** π_θ is pretrained using **Mixed and Annealed Sampling**, which balances targeted practice with broad motion coverage. **Context Representation Learning** extracts in-context latents from the frozen prior's interaction context histories through dynamic prediction and physical neighborhood soft alignment. **Residual Adaptation** learns in-context corrections with both the pretrained prior and context encoder frozen, enabling in-context adaptation without test-time updates. Here, s_t^p denotes the proprioceptive state, m_t the reference motion, a_t^θ the action produced by π_θ, f_i the failure rate, c_i the physical configuration of the i-th environment, E_φ the context encoder, z_i the context representation of c_i, D_ψ the dynamics predictor, π_Δ the in-context residual actor, a_t^Δ the residual action produced by π_Δ, and a_t the final applied action.

3.1 Problem Formulation

We formulate humanoid control under varying physical conditions as a family of Markov decision processes (MDPs) indexed by an environmental configuration $c \in \mathcal{C}$:

$$\mathcal{M}_c = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}_c, \mathcal{R}, \gamma \rangle, \tag{1}$$

where c specifies properties such as mass distribution, friction, and payload, and $\mathcal{T}_c$ is the corresponding transition dynamics. The configuration c is not directly observable at deployment.

Following prior work (Zeng et al., 2026), we use humanoid whole-body motion tracking as the proxy task for BFM pretraining. A conventional domain-randomized BFM $\pi(a_t \mid s_t^p, m_t)$ generates actions a_t from proprioceptive states s_t^p and reference-motion goals m_t using a policy shared across heterogeneous dynamics, which can compromise environment-specific performance. To this end, we introduce **In-Context BFM**, which learns the underlying environmental dynamics from interaction context $\mathcal{H}_t$ and enables adaptive control under varying environmental configurations c.

We formally define the interaction context for humanoid control as the history of proprioceptive states and actions:

$$\mathcal{H}_t \triangleq \left(s_0^p, a_0, s_1^p, a_1, \ldots, s_{t-1}^p, a_{t-1}, s_t^p\right). \tag{2}$$

This history captures the humanoid's responses to past actions and thus provides evidence about the environmental configuration c governing its dynamics. We learn a context encoder E_φ to extract a latent representation of the environmental configuration c from interaction history $\mathcal{H}_t$:

$$z_t = E_\varphi(\mathcal{H}_t) \approx z_c \triangleq g(c), \tag{3}$$

Here, $g : \mathcal{C} \to \mathcal{Z}$ maps each environmental configuration to a dynamics-relevant latent representation z_c, and z_t denotes its context-dependent estimate. The resulting In-Context BFM acts according to $\pi(a_t \mid s_t^p, m_t, z_t)$, enabling adaptation without direct access to privileged parameters.

With the learned encoder E_φ held fixed, our objective is to maximize the expected discounted return over a distribution $p(c)$ of environmental configurations:

$$\max_\pi \mathbb{E}_{\substack{c\sim p(c) \\ \tau\sim p_\pi(\tau|c)}} \left[\sum_{t=0}^{T-1} \gamma^t r_t\right], \tag{4}$$

where r_t denotes the motion-tracking reward and $p_\pi(\tau \mid c)$ is the distribution of length-T rollouts induced by the In-Context BFM π under $\mathcal{T}_c$.

3.2 Learning a General Behavior Prior via MIAS

In-context adaptation across diverse whole-body behaviors requires a sufficiently general behavior prior. We first train the behavior prior π_θ, termed **ECHO-PRE**, in a single stage with MIAS, using motion tracking as the proxy task. MIAS combines failure rewind, uniform exploration, and failure-adaptive sampling to balance broad motion coverage with targeted practice on difficult segments, and gradually reduces sampling concentration to consolidate the learned repertoire.

Adaptive weight. We partition reference motions into temporal bins and estimate their difficulty using empirical failure rates f_i (Luo et al., 2026; Liao et al., 2026). We prioritize bins with observed failures through a length-weighted, temperature-scaled distribution:

$$\widetilde{p}_i \propto \ell_i \exp(f_i/T)\,\mathbf{1}\{f_i > 0\}, \tag{5}$$

where ℓ_i is the number of valid frames in bin i and $T > 0$ controls sampling sharpness. To prevent excessive concentration on individual segments or motions, we cap and redistribute probability mass at both levels, yielding the adaptive distribution p^{A}. Counting rules, the no-failure fallback, and probability caps are detailed in Appendix A.1.

Mixed and annealed sampling. Difficult segments warrant targeted practice without indefinitely restricting motion coverage. MIAS therefore mixes three sampling branches and progressively relaxes the concentration of the adaptive branch.

Mixed sampling. Let $\mathcal{P}_{\mathrm{R}}$, $\mathcal{P}_{\mathrm{U}}$, and $\mathcal{P}_{\mathrm{A}}$ denote the distributions over bins induced by rewind, uniform, and adaptive sampling, respectively. We sample bins from the failure-conditioned mixture:

$$\mathcal{P}_d = d\rho_{\mathrm{R}}\mathcal{P}_{\mathrm{R}} + (1 - d\rho_{\mathrm{R}})\big[\rho_{\mathrm{U}}\mathcal{P}_{\mathrm{U}} + (1-\rho_{\mathrm{U}})\mathcal{P}_{\mathrm{A}}\big], \tag{6}$$

where $d \in \{0,1\}$ indicates whether the preceding trajectory ended in a tracking failure, and $\rho_{\mathrm{R}}, \rho_{\mathrm{U}} \in [0,1]$ specify the rewind probability conditional on failure and the uniform probability conditional on not rewinding, respectively. Rewind restarts the failed motion at the beginning of its failure bin. Uniform and adaptive sampling select a bin uniformly from $\mathcal{I}$ or according to p^{A}, respectively, and then sample a valid frame uniformly within that bin.

Annealed concentration. To gradually broaden practice, we increase the sampling temperature T and decrease the probability cap C using a shared annealing function:

$$\theta_k = \mathcal{A}(k; \theta_{\mathrm{start}}, \theta_{\mathrm{end}}, K_\theta)\,, \qquad \theta \in \{T, C\}, \tag{7}$$

where k is the training iterations and $\mathcal{A}$ performs cosine interpolation with independent horizons K_T and K_C, as specified in Appendix A.1.

3.3 In-Context Adaptation via CRA

CRA first learns a context encoder from the prior's rollouts using dynamics prediction and physically grounded representation objectives, then freezes the context encoder and trains a residual policy to translate context into action corrections.

Interaction context. Let s_t^p denote noisy proprioception and a_t the applied action. At decision time t, the context $\mathcal{H}_t$ contains only completed, valid transitions preceding that decision. We instantiate the context with short- and long-term histories:

$$\mathcal{H}_t = (\mathcal{S}_t, \mathcal{M}_t), \qquad z_t = E_\varphi(\mathcal{H}_t), \tag{8}$$

The short-term history $\mathcal{S}_t$ preserves fine-grained recent responses, while the long-term memory $\mathcal{M}_t$ provides complementary dynamics evidence from earlier interactions across motions. Complete chunks are retained in $\mathcal{M}_t$ across reference-motion changes and episode resets within the same physical session, whereas $\mathcal{S}_t$ is cleared at these boundaries. The physical context remains fixed within a session, and a new session clears all memory.

Predictive in-context dynamics. To encourage the context latent to encode control-relevant dynamics, we jointly train E_φ and a forward predictor D_ψ on rollouts collected by the frozen tracker. Let y_t contain the privileged robot state, foot kinematics, contact forces, and binary contact indicators, and let a_t denote the applied action. Conditioned on the context z_t, the predictor estimates future states over a horizon of K_p steps:

$$\hat{y}_{t+k} = D_\psi(\mathcal{H}_{t+k-1}, a_{t+k-1}; z_t)\,, \tag{9}$$

where $\mathcal{H}_{t+k-1}$ denotes the predictor's rolling input history, initialized from the observed trajectory and updated with recorded states or model predictions during training. The context z_t is held fixed throughout the prediction horizon, and all predictions use the recorded applied target sequence. We minimize the multi-step prediction objective:

$$\mathcal{L}_\mathrm{pred} = \mathbb{E}_{\substack{\tau\sim\mathcal{D} \\ \mathbf{H}\sim\nu(\cdot|\tau)}}\left[\frac{1}{K_\mathrm{p}}\sum_{k=1}^{K_\mathrm{p}} \ell_\mathrm{pred}(\hat{y}_{t+k}, y_{t+k})\right], \tag{10}$$

where $\mathcal{D}$ is the rollout-segment distribution and $\nu(\cdot \mid \tau)$ is the weighted distribution over rolling-history sequences $\mathbf{H} = (\mathcal{H}_t, \ldots, \mathcal{H}_{t+K_\mathrm{p}-1})$.

Physically grounded context geometry. We complement predictive supervision with *physical neighborhood soft alignment* to preserve neighborhoods among physical conditions. Specifically, we align cross-motion context neighborhoods P with physical-parameter neighborhoods Q:

$$\mathcal{L}_\mathrm{soft} = \frac{1}{n}\sum_{i=1}^{n} D_\mathrm{KL}(Q_{i:} \,\|\, P_{i:})\,, \tag{11}$$

where n is the number of eligible randomized systems. More details are in Appendix A.1.

Then, we jointly optimize the context encoder and forward predictor with

$$\mathcal{L}_\mathrm{ctx} = \mathcal{L}_\mathrm{pred} + \mathcal{L}_\mathrm{soft}. \tag{12}$$

Privileged states and physical parameters are used only during training.

In-context residual adaptation. The learned context representation provides evidence about environmental dynamics, which the residual policy translates into environment-specific action corrections. We freeze both the context encoder E_φ and ECHO-PRE π_θ and train a residual actor π_Δ. The predictor is not used in this stage or at deployment. The final action combines the prior's command a_t^θ with a context-conditioned correction a_t^Δ:

$$a_t^\theta = \pi_\theta(s_t^p), \qquad a_t^\Delta = \pi_\Delta(s_t^p, m_t, z_t, a_t^\theta), \qquad a_t = a_t^\theta + a_t^\Delta, \tag{13}$$

where a_t is the final applied action. Each completed valid transition updates the interaction history through Eq. equation 8, closing the interaction–context–control loop.

4 Experiments

4.1 Set up

We use the Unitree G1 for both training and evaluation. We evaluate ECHO extensively in simulation, including tracking fidelity (Section 4.2), in-context adaptation (Section 4.3), and in-context representation analysis (Section 4.4). We further validate ECHO on the real robot in Section 4.5.

To enable fair comparison with existing BFMs while evaluating in-context ability under broader dynamics variations, we design two physical randomization settings. **Necessary Domain Randomization (NDR)** contains the minimum set of DR required for sim-to-real transfer and is designed to remain close to the settings commonly adopted by existing baselines. We pretrain ECHO-PRE only under NDR to acquire the general behavior prior.

We further introduce **Heavy Domain Randomization (HDR)**, which extends NDR with substantially larger dynamics variations, including additional payloads on the wrists and lower legs. Starting from the same frozen ECHO-PRE backbone, we train CRA under NDR and HDR, resulting in

Table 2: **Simulation evaluation of BFMs' motion tracking fidelity.** MIAS improves tracking performance, while CRA further enhances fidelity across both everyday and high-dynamic motions. Results are reported as mean ± sample standard deviation across three evaluation runs with independent action-noise seeds and noise standard deviation $\sigma = 0.01$. Bold and underlined values indicate the best and second-best results within each dataset, respectively.

Method	SR ↑	$E_{\text{g-MPJPE}}$ ↓	E_{MPJPE} ↓	E_{acc} ↓	E_{vel} ↓
AMASS-2000-Random					
SONIC v1.1	0.9633 ± 0.0003	316.651 ± 1.649	30.411 ± 0.015	1.135 ± 0.002	3.770 ± 0.006
HEFT	0.9785 ± 0.0009	261.573 ± 2.352	28.524 ± 0.009	1.062 ± 0.001	3.103 ± 0.004
ECHO-PRE w/o MIAS	0.9720 ± 0.0005	331.235 ± 2.132	32.580 ± 0.018	1.243 ± 0.003	4.170 ± 0.008
ECHO-PRE	**0.9832** ± 0.0012	<u>207.995</u> ± 0.970	<u>26.595</u> ± 0.007	<u>0.932</u> ± 0.003	<u>3.036</u> ± 0.006
ECHO-N	<u>0.9827</u> ± 0.0003	**203.149** ± 0.289	**25.577** ± 0.024	**0.917** ± 0.003	**2.956** ± 0.007
BONES-SEED-100-HighDynamic					
SONIC v1.1	0.5800 ± 0.0100	343.124 ± 4.421	61.791 ± 2.615	4.086 ± 0.031	10.406 ± 0.102
HEFT	0.6200 ± 0.0173	391.094 ± 30.758	59.562 ± 0.135	3.232 ± 0.017	8.229 ± 0.138
ECHO-PRE w/o MIAS	0.4500 ± 0.0100	365.075 ± 5.270	64.580 ± 3.025	4.297 ± 0.042	11.712 ± 0.145
ECHO-PRE	<u>0.7200</u> ± 0.0000	<u>244.772</u> ± 3.143	<u>56.521</u> ± 0.335	<u>2.801</u> ± 0.012	<u>7.507</u> ± 0.034
ECHO-N	**0.7633** ± 0.0115	**235.410** ± 4.140	**55.392** ± 0.415	**2.766** ± 0.005	**7.360** ± 0.017

ECHO-N and **ECHO-H**, respectively. The detailed randomization ranges of Nominal environment, NDR and HDR are summarized in Table 7.

We train ECHO using PPO (Schulman et al., 2017) in Mjlab (Zakka et al., 2026) for behavior pretraining and in-context adaptation, leveraging more than 1,200 hours of filtered high-quality motion data. The composition of the training dataset is illustrated in Figure 5.

4.2 Evaluation of Motion Tracking Fidelity

We evaluate the contributions of MIAS and CRA to motion tracking. To assess MIAS, we compare **ECHO-PRE** with **HEFT** (Liu et al., 2026a), **SONIC v1.1** (Luo et al., 2026), and **ECHO-PRE w/o MIAS**, which uses uniform motion sampling. To assess CRA, we compare **ECHO-N** against its pretrained **ECHO-PRE** backbone. Our evaluation covers two complementary benchmarks: **AMASS-2000-Random** tests generalization to everyday motions from AMASS, which is excluded from ECHO's training data, while **BONES-SEED-100-HighDynamic** tests challenging motions within the training distribution. All evaluations use nominal dynamics without domain randomization. We report success rate (SR) and motion-tracking errors, with dataset construction and metric definitions detailed in Appendix A.3.

As shown in Table 2, MIAS improves tracking on both benchmarks relative to uniform sampling, and ECHO-PRE compares favorably with the external baselines. CRA further reduces tracking errors, showing that interaction context improves execution precision even under nominal dynamics.

4.3 Evaluating the ability of using context

In this section, we compare ECHO with representative adaptation methods for legged robot control. We consider three baselines. **RMA** (Kumar et al., 2021) learns a privileged teacher conditioned on ground-truth dynamics parameters and subsequently distills its adaptation capability into a student policy; we report both the **RMA-Teacher** and **RMA-Student** for comparison. **Any2Track** Zhang et al. (2025b) learns a latent dynamics representation from interaction history and regularizes the representation with a world-model prediction objective. We further design a **Vanilla** baseline, which uses the same residual-policy architecture but receives only the observations available to ECHO-PRE, without any latent dynamics representation.

For a fair comparison, all methods share the same frozen ECHO-PRE backbone and residual-policy architecture, and are trained under the same HDR setting. We evaluate all methods on AMASS-2000-Random under both the nominal environment and the HDR environment. The evaluation metrics are identical to those used in Section 4.2.

Table 3: **Simulation evaluation of BFMs' in-context ability.** ECHO-H effectively exploits interaction context to improve tracking under heavily randomized dynamics (HDR). Simulation evaluation of BFMs' in-context ability. Bold and underlined values indicate the best and second-best results within each environment, respectively.

Method	SR ↑	$E_{\text{g-MPJPE}}$ ↓	E_{MPJPE} ↓	E_{acc} ↓	E_{vel} ↓
Nominal Environment					
Vanilla	0.8892 ± 0.0003	194.541 ± 1.585	27.286 ± 0.016	0.881 ± 0.008	2.887 ± 0.006
Any2Track	0.8743 ± 0.0008	193.623 ± 0.738	27.846 ± 0.002	0.873 ± 0.008	2.883 ± 0.005
RMA-Teacher	**0.9798** ± 0.0003	**189.908** ± 0.220	<u>24.062</u> ± 0.010	<u>0.849</u> ± 0.007	<u>2.772</u> ± 0.004
RMA-Student	<u>0.9763</u> ± 0.0008	190.329 ± 0.590	**24.026** ± 0.007	0.851 ± 0.008	2.774 ± 0.005
ECHO-H	0.9748 ± 0.0010	<u>190.088</u> ± 0.718	24.496 ± 0.014	**0.846** ± 0.008	**2.760** ± 0.006
HDR Environment					
Vanilla	0.8632 ± 0.0008	230.001 ± 0.833	31.541 ± 0.018	1.088 ± 0.003	3.701 ± 0.004
Any2Track	0.8455 ± 0.0009	227.549 ± 0.999	31.939 ± 0.012	1.087 ± 0.003	3.709 ± 0.003
RMA-Teacher	<u>0.8827</u> ± 0.0021	<u>216.826</u> ± 1.462	**30.063** ± 0.017	<u>1.052</u> ± 0.002	<u>3.591</u> ± 0.002
RMA-Student	0.8700 ± 0.0009	224.568 ± 0.894	30.542 ± 0.014	1.065 ± 0.002	3.648 ± 0.001
ECHO-H	**0.8922** ± 0.0016	**210.351** ± 1.017	<u>30.463</u> ± 0.008	**1.039** ± 0.003	**3.530** ± 0.003

As shown in Table 3, ECHO-H achieves performance comparable to RMA-Teacher in the nominal environment, without requiring privileged dynamics parameters at inference. Under HDR, ECHO-H achieves the best results on four of the five metrics, improving SR by 0.95 percentage points and reducing global MPJPE by 3.0% relative to RMA-Teacher. ECHO-H also outperforms RMA-Student across all HDR metrics and consistently improves over Vanilla in both environments. In contrast, the Any2Track baseline provides no consistent gains over Vanilla, suggesting that introducing a history-derived latent alone does not guarantee effective adaptation in this setting. Together, these results suggest that ECHO learns control-relevant representations of environmental dynamics that the residual policy can effectively exploit to improve tracking under substantial dynamics variation.

4.4 Analyzing the learned context representations

We examine whether CRA learns environment representations that remain consistent across motions while retaining physically meaningful information. We collect trajectories by rolling out ECHO-PRE across 40 LAFAN motions in 1,024 HDR environments and evaluate all variants on 325,193 shared valid history windows. We evaluate environment identification, cross-motion retrieval, separation, and linear readout of payload mass and friction. See Appendix A.4 for details.

As reported in Table 4, ECHO achieves 97.95% environment identification accuracy and 94.96% cross-motion retrieval mAP, with linear-readout R^2 scores of 0.846 for payload mass and 0.926 for friction. These results indicate that the learned representations capture environment information consistently across motions while keeping physical properties linearly accessible. The t-SNE visualization (Maaten & Hinton, 2008) in Fig. 3 provides qualitative support, showing clearer environment-wise grouping than our Any2Track implementation (Zhang et al., 2025b) on 16 randomly selected environments. Additional visualizations on AMASS are provided in Appendix Fig. 8.

The loss ablations reveal complementary roles of soft alignment and the predictor objective. Removing physical-neighborhood soft alignment reduces cross-motion mAP from 94.96% to 92.21% and the silhouette score from 0.4718 to 0.4369, suggesting weaker cross-motion consistency and environment separation. In contrast, removing the predictor leaves retrieval and separation scores nearly unchanged, but reduces payload mass R^2 from 0.846 to 0.481 and friction R^2 from 0.926 to -0.014. This contrast highlights the predictor's contribution to making physical properties linearly accessible, beyond merely distinguishing environments.

Scaling interaction context. Our context encoder supports longer interaction histories by aggregating additional long-term chunks into a fixed-size memory representation. We fix the short-term history at 1 second (50 steps) and evaluate total context budgets of 2, 4, 6, 7, 8,and 9 seconds. Each variant is trained separately for 5,000 updates using the prediction loss and soft-alignment loss. Evaluation uses 325,629 shared, unaveraged history windows from 40 LAFAN (Harvey et al., 2020)

Table 4: **Evaluation of context representations on LAFAN-40.** Results are mean ± sample standard deviation over five evaluation seeds with independently sampled window starts and observation noise. Encoder checkpoints, the 1,024 physical environments, and probe splits are fixed; all methods share trajectories and valid windows within each seed. Mass denotes added payloads. Bold and underlined indicate the best and second-best results, ranked before rounding.

Method	Env. Acc. (%) ↑	mAP (%) ↑	Silhouette ↑	Mass R^2 ↑	Friction R^2 ↑
w/o soft alignment	97.125 ± 0.636	91.282 ± 1.233	0.4337 ± 0.0084	<u>0.7869</u> ± 0.0061	<u>0.9069</u> ± 0.0078
w/o predictor	<u>97.141</u> ± 0.743	<u>93.761</u> ± 1.482	<u>0.4555</u> ± 0.0143	0.4777 ± 0.0025	-0.0132 ± 0.0003
ECHO	**97.478** ± 0.762	**93.872** ± 1.450	**0.4642** ± 0.0105	**0.8357** ± 0.0080	**0.9208** ± 0.0088

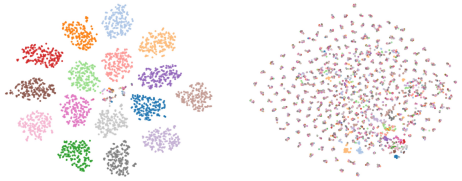

Figure 3: **Context representation comparison.** t-SNE visualizations of ECHO (left) and Any2Track (right), with colors indicating environments. ECHO shows clearer environment-wise grouping across motions.

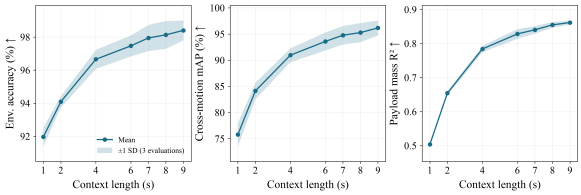

Figure 4: **Context-length scaling of environment representations.** Linear environment identification accuracy, cross-motion retrieval mAP, and payload-mass linear readout R^2 on LAFAN-40 (left to right). The points and shaded bands denote the mean and sample standard deviation over three evaluations.

motions. Figure 4 shows consistent improvements with longer context: from 50 to 350 steps, linear environment identification accuracy increases from 91.50% to 97.95%, and cross-motion retrieval mAP from 73.61% to 94.95%. Payload mass R^2 increases from 0.498 to 0.846, showing improved linear readability of physical properties. These results demonstrate that our method can scale its interaction context to learn more informative environment representations with better cross-motion consistency. Complete results and protocol details are provided in Appendix A.4 (Table 8).

4.5 Real-world experiments

We demonstrate ECHO's diverse whole-body behaviors through real-time teleoperation, including crawling, highly dynamic motions, and terrain traversal (Figure 1, Figure 6, and the accompanying video). As shown in Figure 7, in-context adaptation enables ECHO-H to learn context-dependent control under HDR and maintain precise tracking with added payloads. Even with 10 kg of payloads distributed across its hands and legs (three 2 kg loads and one 4 kg load), ECHO-H can still stand on one leg, squat, and bend forward. For quantitative evaluation, we execute four reference motions and compute MPJPE-L from joint positions recorded during deployment. We compare ECHO-N with ECHO-PRE, HEFT, and SONIC v1.1 without additional payloads, and ECHO-H with Vanilla and RMA-Student under a 2 kg load on each lower leg. As reported in Table 9, ECHO-N and ECHO-H achieve the lowest tracking errors across all four motions in their respective settings, supporting the effectiveness of context-conditioned adaptation for real-world motion tracking.

5 Conclusion

We presented ECHO, an in-context behavior foundation model that separates general behavioral capability from dynamics-specific execution. MIAS acquires a broad whole-body repertoire, while CRA learns physical dynamics from interaction context to guide residual corrections to the frozen behavior prior. Experiments and real-world deployment on the Unitree G1 demonstrate improved tracking, whole-body teleoperation, and adaptation to added payloads. These results show that interaction context enables a shared behavior prior to adapt to its physical environment, combining behavioral breadth with dynamics-specific precision without test-time parameter updates.

AI USE STATEMENT

In this work, we used generative AI tools, to assist with translation and to obtain feedback on research methodology, experimental design, and the interpretation of experimental results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, or formulate mathematical claims. Additionally, we used these tools for manuscript drafting and organization, language editing, and LaTeX formatting. The authors reviewed and revised all AI-assisted content, checked references and related-work descriptions against the original publications, and verified technical descriptions and result interpretations against the implementation and experimental records. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Yoshua Bengio, Ian Goodfellow, Aaron Courville, et al. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

Fu Chen, Xin Ding, Bingjia Huang, Xiangyu Li, Mingju Wang, Jiawei He, Kun Li, Wei Sun, Yunxin Liu, Hao Wu, et al. Zeva: In-context causal learning for generalizable embodied manipulation. *arXiv preprint arXiv:2608.30880*, 2026a.

Maiyue Chen, Kaihui Wang, Bo Zhang, Xihan Ma, Zhiyuan Yang, Yi Ren, Qijun Huang, Zihao Zhu, Yucheng Wang, and Zhizhong Su. Holomotion-1 technical report. *arXiv preprint arXiv:2605.15336*, 2026b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xiao Chen, Weishuai Zeng, Xiaojie Niu, Zirui Wang, Jianan Li, Huayi Wang, Furui Xu, Jiahe Chen, Weixiang Zhong, Lihe Ding, et al. Reactivebfm: Reactive closed-loop motion planning towards universal humanoid whole-body control. *arXiv preprint arXiv:2606.30362*, 2026c.

Ziyang Cheng, Tianshu Tang, Jinxin Lan, Xinze Chen, Yuhan Gong, Zhichao Liu, Changzhong Wu, Yahao Mao, Zongyan Deng, Mingxuan Ma, et al. Gigabrain-wbc-0.5: A behavior world model for robust whole-body control with environment interaction. *arXiv preprint arXiv:2608.18234*, 2026.

ChingMu. MotionDecode: Chingmu 1000-hour embodied motion dataset. Dataset, 2026. URL https://huggingface.co/datasets/CMRobot/MotionDecode.

Xinyang Gu, Yen-Jen Wang, Xiang Zhu, Chengming Shi, Yanjiang Guo, Yichen Liu, and Jianyu Chen. Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning. *arXiv preprint arXiv:2408.14472*, 2024.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion inbetweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 international conference on robotics and automation (ICRA)*, pp. 6023–6029. IEEE, 2019.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, et al. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 79697–79725, 2026.

Zhongyu Li, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, 44(5):840–888, 2025.

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *Science Robotics*, 11(117):eadx8924, 2026.

Chenxin Liu, Qingzhou Lu, Guangxiao Yang, Xuanyang Shi, Chenghan Yang, Yanjiang Guo, and Jianyu Chen. Heft: Heavy-payload full-size humanoid teleoperation with privileged motion guidance and windowed payload curriculum. *arXiv preprint arXiv:2607.02332*, 2026a.

Dairu Liu, Zekun Qi, Jiayu Zeng, Yu Guan, Chenghuai Lin, Xuchuan Chen, Xinqiang Yu, Wenyao Zhang, He Wang, and Li Yi. Humantracker: Towards comprehensive and human-aligned motion tracking benchmark. In *European Conference on Computer Vision*, pp. 133–152. Springer, 2026b.

Min Liu, Deepak Pathak, and Ananye Agarwal. Locoformer: Generalist locomotion via long-context adaptation. *arXiv preprint arXiv:2509.23745*, 2025.

Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

Yubiao Ma, Han Yu, Kai Guo, Changtai Lv, Zhengquan Mao, Boyang Xing, Xuemei Ren, and Dongdong Zheng. Extreme-rgmt: Continual learning of highly dynamic skills for robust generalist humanoid control. *arXiv preprint arXiv:2607.20110*, 2026.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL `https://nlp.stanford.edu/IR-book/`.

Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022.

Bhairav Mehta, Manfred Diaz, Florian Golemo, Christopher J Pal, and Liam Paull. Active domain randomization. In *Conference on robot learning*, pp. 1162–1176. PMLR, 2020.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *European conference on computer vision*, pp. 283–299. Springer, 2018.

Zekun Qi, Xuchuan Chen, Dairu Liu, Chenghuai Lin, Yunrui Lian, Sikai Liang, Zhikai Zhang, Yu Guan, Jilong Wang, Wenyao Zhang, et al. Humanoid-gpt: Scaling data and structure for zero-shot motion tracking. *arXiv preprint arXiv:2606.03985*, 2026.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8): 913–929, 2017. doi: 10.1111/ecog.02881.

RoboParty Lab Team. Mimiclite: Efficient and effective general humanoid motion tracking. `https://github.com/EGalahad/mimic-lite`, 2026. Technical report: `https://github.com/Roboparty/MimicLite/blob/main/mimic-lite.pdf`.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

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

Rutav Shah, Shuijing Liu, Qi Wang, Zhenyu Jiang, Sateesh Kumar, Mingyo Seo, Roberto Martín-Martín, and Yuke Zhu. Mimicdroid: In-context learning for humanoid robot manipulation from human play videos. *arXiv preprint arXiv:2509.09769*, 2025.

Bones Studio. Bones-seed: Skeletal everyday embodiment dataset, 2026. *URL https://huggingface.co/datasets/bones-studio/seed*, 142.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Puyue Wang, Jiawei Hu, Yan Gao, Junyan Wang, Yu Zhang, Gillian Dobbie, Tao Gu, Wafa Johal, Ting Dang, and Hong Jia. Hord: Robust humanoid control via history-conditioned reinforcement learning and online distillation. *arXiv preprint arXiv:2602.04412*, 2026a.

Xingyi Wang, Chenyun Zhang, Weiji Xie, Chao Yu, Wei Song, Chenjia Bai, and Shiqiang Zhu. Halo: Closing sim-to-real gap for heavy-loaded humanoid agile motion skills via differentiable simulation. *arXiv preprint arXiv:2603.15084*, 2026b.

Yunshen Wang, Shaohang Zhu, Peiyuan Zhi, Yuhan Li, Jiaxin Li, Yong-Lu Li, Yuchen Xiao, Xingxing Wang, Baoxiong Jia, and Siyuan Huang. Omnixtreme: Breaking the generality barrier in high-dynamic humanoid control. *arXiv preprint arXiv:2602.23843*, 2026c.

Zepeng Wang, Jiangxing Wang, Shiqing Yao, Yu Zhang, Ziluo Ding, Ming Yang, Yuxuan Wang, Haobin Jiang, Chao Ma, Xiaochuan Shi, et al. General humanoid whole-body control via pretraining and fast adaptation. *arXiv preprint arXiv:2602.11929*, 2026d.

Zifan Wang, Yizhao Li, Teli Ma, Qiang Zhang, Yudong Fan, Hao Xu, Shuo Yang, and Junwei Liang. Perceptive behavior foundation model: Adapting human motion priors to robot-centric terrain. *arXiv preprint arXiv:2606.08059*, 2026e.

Angchen Xie, Nikhil Sobanbabu, Ishayu Shikhare, Alan Wang, Max Simchowitz, and Guanya Shi. Fada: Few-shot domain adaptation via dynamics alignment for humanoid control. *arXiv preprint arXiv:2606.28476*, 2026.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pp. 24631–24645. PMLR, 2022.

Wenhao Yu, Jie Tan, C Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. *arXiv preprint arXiv:1702.02453*, 2017.

Kevin Zakka, Qiayuan Liao, Brent Yi, Louis Le Lay, Koushil Sreenath, and Pieter Abbeel. mjlab: A lightweight framework for gpu-accelerated robot learning. *arXiv preprint arXiv:2601.22074*, 2026.

Weishuai Zeng, Shunlin Lu, Kangning Yin, Xiaojie Niu, Minyue Dai, Jingbo Wang, and Jiangmiao Pang. Behavior foundation model for humanoid robots. *arXiv preprint arXiv:2509.13780*, 2025.

Weishuai Zeng, Kangning Yin, Xiaojie Niu, Shunlin Lu, Weixiang Zhong, Jiahe Chen, Feiyu Jia, Xiao Chen, Zirui Wang, Furui Xu, et al. Scaling behavior foundation model for humanoid robots. *arXiv preprint arXiv:2607.15163*, 2026.

Heng Zhang, Wei-Hsing Huang, Gokhan Solak, and Arash Ajoudani. Omnivic: A self-improving variable impedance controller with vision-language in-context learning for safe robotic manipulation. *arXiv preprint arXiv:2510.17150*, 2025a.

Zhikai Zhang, Jun Guo, Chao Chen, Jilong Wang, Chenghuai Lin, Yunrui Lian, Han Xue, Zhenrong Wang, Maoqi Liu, Jiangran Lyu, et al. Track any motions under any disturbances. *arXiv preprint arXiv:2509.13833*, 2025b.

A Appendix

A.1 Details of Methods

Adaptive sampling. Let $\mathcal{I}$ denote the set of valid temporal bins, where $i = (m, b)$ identifies bin b of motion m. For each bin, we maintain completed-visit and failure counts, V_i and F_i, and estimate its failure rate as $f_i = F_i/V_i$ for $V_i > 0$. A visit is completed when a trajectory leaves the bin or terminates. We compute the preliminary sampling distribution as

$$s_i = \ell_i \exp\left(\frac{f_i - f_{\max}}{T}\right) \mathbf{1}\{f_i > 0\},$$
$$\widetilde{p}_i = \begin{cases} \dfrac{s_i}{\sum_{j\in\mathcal{I}} s_j}, & \sum_{j\in\mathcal{I}} s_j > 0, \\ \dfrac{\ell_i}{\sum_{j\in\mathcal{I}} \ell_j}, & \text{otherwise}, \end{cases} \tag{14}$$

where $f_{\max} = \max_{j\in\mathcal{I}} f_j$ is subtracted for numerical stability. Length weighting accounts for partially filled bins and provides a length-proportional fallback when no failures have been observed. We then cap and redistribute probability mass to obtain p^{A} satisfying

$$\sum_{i\in\mathcal{I}} p_i^{\mathrm{A}} = 1, \qquad 0 \le p_i^{\mathrm{A}} \le \frac{C}{N}, \qquad \sum_{i\in\mathcal{I}_m} p_i^{\mathrm{A}} \le \frac{C}{M} \quad \forall m, \tag{15}$$

where $N = |\mathcal{I}|$, M is the number of motions, and $\mathcal{I}_m$ contains the bins of motion m. The multiplier $C \ge 1$ is chosen so that the constraints are jointly feasible, and redistribution preserves relative weights wherever capacity permits.

Mixed sampling. In practice, we set $\rho_{\mathrm{R}} = 1/3$ and $\rho_{\mathrm{U}} = 1/2$ in Eq. equation 6, yielding

$$\mathcal{P}_d = \frac{d}{3}\mathcal{P}_{\mathrm{R}} + \frac{3-d}{6}\big(\mathcal{P}_{\mathrm{U}} + \mathcal{P}_{\mathrm{A}}\big). \tag{16}$$

After a tracking failure ($d = 1$), all three components receive equal weight $1/3$. After a non-failure timeout or normal motion completion ($d = 0$), rewind is disabled, and uniform and adaptive sampling each receive weight $1/2$. A tracking failure coinciding with a timeout is assigned $d = 1$.

The rewind component $\mathcal{P}_{\mathrm{R}}$ selects the failure bin of the preceding motion and restarts at the beginning of that bin. For each valid bin $i \in \mathcal{I}$, the uniform and adaptive components satisfy $\mathcal{P}_{\mathrm{U}}(i) = 1/|\mathcal{I}|$ and $\mathcal{P}_{\mathrm{A}}(i) = p_i^{\mathrm{A}}$, respectively. Both components then sample a valid frame τ uniformly within the selected bin. Uniform sampling initializes at τ, whereas adaptive sampling additionally applies a random backward offset to include the lead-up to difficult segments:

$$t_0 = \max\{0, \tau - u\}, \qquad u \sim \mathrm{Uniform}\{0, \dots, W-1\}, \tag{17}$$

where τ and t_0 are frame indices within the selected motion, and W is a positive integer specifying the backward-offset window in frames.

Annealing schedule. The annealing function in Eq. equation 7 interpolates from an initial value a to a final value b over a horizon of K iterations:

$$\mathcal{A}(k; a, b, K) = a + \frac{b-a}{2}\left[1 - \cos(\pi \min\{k/K, 1\})\right], \tag{18}$$

where $k \geq 0$ and $K > 0$. For the sampling temperature, we use $(a, b, K) = (T_{\text{start}}, T_{\text{end}}, K_T)$, with $T_{\text{end}} > T_{\text{start}} > 0$. For the cap multiplier, we use $(a, b, K) = (C_{\text{start}}, C_{\text{end}}, K_C)$, with $C_{\text{start}} > C_{\text{end}} \geq 1$. Each schedule reaches its final value at the corresponding horizon and remains constant thereafter. All scheduled cap values are chosen to keep the constraints in Eq. equation 15 jointly feasible.

Prediction loss. The per-step prediction loss is

$$\ell_{\text{pred}}(\hat{y}, y) = \sum_{g \in \mathcal{G}} \alpha_g \, \text{Huber}(e_g(\hat{y}, y)) + \alpha_{\text{bin}} \, \text{BCE}(\hat{y}^{\text{bin}}, y^{\text{bin}}). \tag{19}$$

Here, $\mathcal{G}$ comprises root position and orientation, linear and angular velocity, joint position and velocity, foot kinematics, and contact force. Each e_g is a scaled component-wise error, with rotation-vector errors for orientation and Huber losses (Huber, 1992) averaged over coordinates. The binary cross entropy (Bengio et al., 2017) term supervises contact occurrence.

Physical neighborhood soft alignment. For n eligible randomized systems in a minibatch, let ξ_i denote the range-normalized, group-balanced physical-parameter vector of system i, with $d_{ij} = \|\xi_i - \xi_j\|_2$. Let $\bar{z}_i$ and $\bar{z}'_i$ denote the mean context embeddings from its current and archived histories, respectively. A *current history* contains the completed state–action–next-state transitions available at the corresponding training sample's time; "current" is relative to that sample and does not imply recently collected data. An *archived history* is an earlier, previously stored interaction window from the same environment instance and the same physical session, during which the physical parameters remain unchanged. The paired current and archived windows correspond to different reference motions and do not overlap in time. Only systems with an eligible history pair are included in the neighborhood loss. Our neighborhood-distribution matching is related to probabilistic knowledge transfer (Passalis & Tefas, 2018), but uses physical-parameter distances to supervise cross-motion context similarities. We construct the physical and latent neighborhood distributions as

$$\begin{aligned} Q_{ij} &= \text{softmax}_j\left(-(\log 2)\left(\frac{d_{ij}}{h_c}\right)^2\right), \\ P_{ij} &= \text{softmax}_j\left(\frac{\bar{z}_i^\top \bar{z}'_j}{\tau_z}\right), \end{aligned} \tag{20}$$

where softmax normalizes each row over all n systems, $h_c > 0$ controls the physical-neighborhood bandwidth, and $\tau_z > 0$ is the latent similarity temperature. The neighborhood loss in Eq. equation 11 is the average row-wise KL divergence:

$$\mathcal{L}_{\text{soft}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij} \log \frac{Q_{ij}}{P_{ij}}. \tag{21}$$

Matching neighborhoods across different motions encourages the learned relationships to reflect physical conditions rather than motion identity.

Joint optimization. We jointly optimize the context encoder E_φ and forward predictor D_ψ using the total objective in Eq. equation 12. Each loss term is averaged over its eligible samples. Privileged states and physical parameters are used only during training; the context encoder receives interaction history alone.

A.2 Details of training

Datasets. As summarized in Figure 5, our filtered training corpus contains 220,560 motion sequences and approximately 222.46 million frames, corresponding to 1,235.88 hours at 50 frames

per second. It comprises MotionDecode (ChingMu, 2026) (953.17 hours), BONES-SEED (Studio) (260.85 hours), 100STYLE (Mason et al., 2022) (18.63 hours), and LAFAN1 (Harvey et al., 2020) (2.52 hours). The corpus additionally includes two augmentation subsets, *Augmentation Data Flat* and *Augmentation Data Terrain*, containing 75,158 and 51,765 frames, respectively. All reported quantities refer to the data retained after filtering, rather than the complete source datasets. AMASS Mahmood et al. (2019) is excluded from training and used to evaluate generalization to unseen motions. Tables 5 and 6 summarize the rewards and observations used to train ECHO-PRE. Tracking terms use exponential error shaping as in HEFT (Liu et al., 2026a).

Reward and Observation Specifications Tables 5 and 6 summarize the rewards and observations used to train ECHO-PRE. Tracking terms use exponential error shaping as in HEFT (Liu et al., 2026a).

Domain randomization settings. Table 7 summarizes the nominal configuration and the two domain randomization settings, NDR and HDR. The nominal environment uses the baseline physical parameters with a foot friction coefficient of 0.6, no additional payloads, and no external force perturbations. NDR follows the original tracker configuration, randomizing torso mass and center of mass (COM), foot friction, controller gains, and joint armature, while applying intermittent external forces. HDR retains all NDR ranges and additionally introduces four-limb payload randomization, with $[0,\ 2.5]$ kg per hand and $[0,\ 4.0]$ kg per shin (up to 13 kg in total), payload COM offsets of ± 5 cm per axis, and inertia recomputed to account for the added payloads.

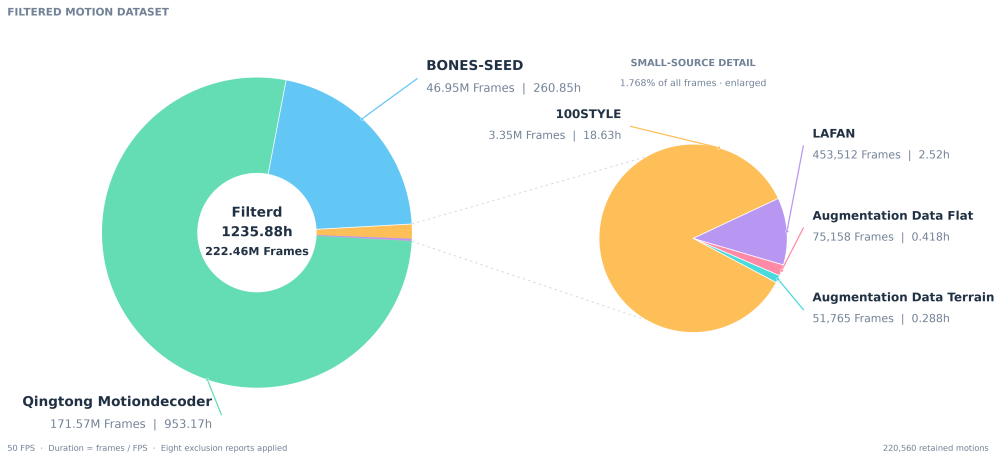

Figure 5: **Composition of the filtered training corpus.** The corpus comprises 1,235.88 hours of motion data (222.46 million frames). The left chart shows source proportions by frame count, while the right chart provides an enlarged breakdown of the smaller sources.

A.3 EVALUATION OF MOTION TRACKING FIDELITY

Benchmark construction. We construct two benchmarks to assess complementary aspects of motion tracking. **AMASS-2000-Random** contains 2,000 randomly sampled motions from AMASS (Mahmood et al., 2019). Since AMASS is excluded from ECHO's training data, this benchmark evaluates ECHO's out-of-distribution generalization to diverse everyday behaviors. **BONES-SEED-100-HighDynamic** contains 100 highly dynamic motions sampled from BONES-SEED (Studio): 33 ground-interaction motions, 33 extreme motions, and 34 dance motions. These motions belong to the training distribution and assess the policy's mastery of challenging behaviors.

Evaluation metrics and simulator. Following Luo et al. (2023; 2026), we report success rate (SR) and four discrepancy metrics between executed and reference trajectories: global MPJPE ($E_{\text{g-MPJPE}}$, mm), root-relative MPJPE (E_{MPJPE}, mm), acceleration error (E_{acc}, mm/frame^2), and velocity error (E_{vel}, mm/frame). Higher SR and lower tracking errors indicate better performance.

Table 5: **Reward terms used for ECHO-PRE training.**

Reward term	**Weight** w_k	**Unweighted formula** r_k
root_pos_tracking	1.5	$\exp[-\|p-\hat{p}\|_2/0.3]$
root_rot_tracking	1.0	$\exp[-d(R,\hat{R})/0.4]$
root_vel_tracking	1.0	$\exp[-\|R^\top v-\hat{R}^\top\hat{v}\|_2/1.0]$
root_ang_vel_tracking	1.0	$\exp[-\|R^\top\omega-\hat{R}^\top\hat{\omega}\|_2/3.0]$
keypoint_pos_tracking	2.0	$\exp[-\sum_i \|x_i^b-\hat{x}_i^b\|_2/(0.3K)]$
keypoint_vel_tracking	1.0	$\exp[-\sum_i \|v_i^b-\hat{v}_i^b\|_2/(1.0K)]$
keypoint_rot_tracking	2.0	$\exp[-\sum_i d(R_i^b,\hat{R}_i^b)/(0.4K)]$
keypoint_angvel_tracking	1.0	$\exp[-\sum_i \|\omega_i^b-\hat{\omega}_i^b\|_2/(3.0K)]$
joint_pos_tracking	1.0	$\exp[-\|q-\hat{q}\|_1/(0.5n)]$
joint_vel_tracking	0.5	$\exp[-\|\dot{q}-\hat{\dot{q}}\|_1/(3.0n)]$
survival	3.0	1
joint_vel_l2	3.5×10^{-4}	$-\sum_j \bar{\dot{q}}_j^2$
waist_joint_acc_l2	10^{-5}	$-\sum_{j\in\mathcal{W}}\langle\ddot{q}_j^2\rangle$
joint_power_l1	2×10^{-5}	$-\sum_j\langle\|\tau_j\dot{q}_j\|\rangle$
knee_joint_power_l1	5×10^{-5}	$-\sum_{j\in\mathcal{K}}\langle\|\tau_j\dot{q}_j\|\rangle$
action_rate_l2	0.3	$-\|a_t-a_{t-1}\|_2^2$
waist_action_rate_l2	-5.0	$\sum_{j\in\mathcal{W}}(a_{t,j}-a_{t-1,j})^2$
feet_air_time_ref	2.5	$\sum_f c_{f,t}(1-c_{f,t-1})\min(T_{f,t}-0.25,0)$
feet_air_time_ref_dense	1.0	$-\frac{1}{2}\sum_f \begin{cases}1, & c_f\neq\hat{c}_f,\\ 1-C_a(h_f^-), & c_f=\hat{c}_f=0,\\ C_c(h_f^+), & c_f=\hat{c}_f=1\end{cases}$
joint_pos_limits	1.5	$-10\sum_j\big([q_j^- - q_j]_+ + [q_j-q_j^+]_+\big)$
joint_torque_limits	0.015	$-\sum_j[\|\tau_j\|/(0.75\tau_j^{\max})-1]_+$

The per-step reward is $r_t=\Delta t\sum_k w_k r_k(t)$, with $\Delta t=0.02\,$s. Hats denote reference states; p,R,v,ω denote the root state, q joint positions, τ joint torques, and a raw actions. d is the relative rotation angle, $K=13$, $n=29$, and $\mathcal{W},\mathcal{K}$ are the waist and knee joint sets. Keypoint quantities with superscript b are root-relative and expressed in the respective root frames. $\bar{\dot{q}}$ averages the last two physics substeps; $\langle\cdot\rangle$ averages all four. For the two feet, $c_f,\hat{c}_f$ are contact indicators and $T_{f,t}=(1-c_{f,t-1})T_{f,t-1}+\Delta t(1-2|c_{f,t}-\hat{c}_{f,t}|)$, initialized at zero. h_f^-,h_f^+ are the minimum and maximum world-frame ankle/toe heights; $C_a(h)=\mathrm{clip}_{[0,1]}((h-0.035)/0.12)$ and $C_c(h)=\mathrm{clip}_{[0,1]}((h-0.035)/0.09)$. Finally, $[x]_+=\max(x,0)$, $q_j^\pm$ are the 90% soft position limits, and $\tau_j^{\max}$ is the actuator torque limit.

All policies are evaluated using the CPU implementation of MuJoCo (Todorov et al., 2012) under nominal dynamics, without domain randomization during evaluation.

A.4 Context Representation Evaluation

Sampling and splits. Each environment–motion pair contributes up to eight independent sessions, with one latent extracted after 500 warm-up steps. ECHO variants are evaluated after 5,000 training iterations; the full encoder produces a 64-dimensional context from at most 350 completed interactions. Following grouped evaluation (Roberts et al., 2017), we split motions into five folds, keeping each motion's windows together. For classification, three folds fit the probe and one selects hyperparameters; the probe is then refitted on all four non-test folds. Retrieval uses these four folds as the gallery. Physical readout additionally splits environments into five folds, yielding 25 tests that exclude both test environments and test motions from probe fitting and validation. Preprocessing uses training data only. LAFAN motions belong to the encoder's pretraining corpus.

Metrics. We report frozen-encoder linear classification accuracy (Chen et al., 2020), full-gallery mAP (Manning et al., 2008), and the Silhouette score (Rousseeuw, 1987) using environment identities. Retrieval treats windows from the same environment as positives; both retrieval and Silhouette use Euclidean distances between original latent vectors without additional normalization. Ridge probes (Hoerl & Kennard, 1970) predict payload mass and friction. Probe scores pool out-of-fold

Table 6: **Observations used for ECHO-PRE training.**

Signal	Definition
Actor observations	
Proprioceptive history	Recent joint states, IMU signals, actions, and torques.
Estimated root height	Root height predicted by the state estimator.
Estimated foot contacts	Predicted foot contact probabilities.
Key-body state	Current key-body poses and velocities in the root frame.
Reference root/joint features	Root and joint targets derived from the denoised reference.
Reference key-body state	Key-body poses and velocities from the denoised reference.
State tracking errors	Errors in joint states, gravity, angular velocity, and height.
Key-body tracking errors	Key-body pose and velocity differences from the reference.
Critic observations	
Noisy reference features	Noisy motion targets and their tracking errors.
Noisy proprioceptive history	Joint states and IMU signals with observation noise.
Action history	Previous policy mean actions.
Privileged reference features	Clean reference root and joint motion targets.
Privileged key-body state	Key-body poses and velocities from the simulator.
Reference key-body pose/errors	Reference key-body poses and corresponding tracking errors.
Privileged proprioceptive history	True joint states, gravity, and root velocities.
Applied actions/torques	Applied joint position targets and actuator torques.
Foot contact state	Contact forces, states, durations, and reference contact flags.

Table 7: **Domain randomization ranges.** Nominal denotes the clean environment used for evaluation. NDR follows the original tracker configuration, while HDR extends NDR with four-limb payload randomization.

Parameter	**Nominal**	NDR	HDR
Torso COM offset (cm/axis)	$\mathbf{0}$	± 7.5	± 7.5
Torso mass offset (kg)	$\mathbf{0}$	$[-1,\ 1]$	$[-1,\ 1]$
Foot friction coefficient	$\mathbf{0.6}$	$[0.3,\ 2.0]$	$[0.3,\ 2.0]$
Upper-limb K_p, K_d scale	$\mathbf{1.0}$	$[0.9,\ 1.1]$	$[0.9,\ 1.1]$
Waist/lower-limb K_p, K_d scale	$\mathbf{1.0}$	$[0.8,\ 1.2]$	$[0.8,\ 1.2]$
Joint armature scale	$\mathbf{1.0}$	$[0.8,\ 1.2]$	$[0.8,\ 1.2]$
External force (N/axis)	$\mathbf{0}$	± 10	± 10
Force duration (s)	–	$[0.3,\ 0.5]$	$[0.3,\ 0.5]$
Force interval (s)	–	$[3,\ 6]$	$[3,\ 6]$
Payload per hand (kg)	$\mathbf{0}$	0	$[0,\ 2.5]$
Payload per shin (kg)	$\mathbf{0}$	0	$[0,\ 4.0]$
Payload COM offset (cm/axis)	–	–	± 5
Total payload mass (kg)	$\mathbf{0}$	0	$[0,\ 13]$
Payload-induced inertia	Unchanged	Unchanged	Recomputed

predictions across valid windows; mass R^2 averages the four limb-wise scores, while friction R^2 evaluates a single coefficient.

Visualization. Both panels use the same windows and t-SNE settings: perplexity 5, PCA initialization and 1,500 iterations. The Any2Track comparison is qualitative because inputs, capacity, history length, and training setup differ. To further evaluate out-of-distribution generalization, we visualize the learned context representations on the AMASS dataset using t-SNE.

A.5 Real world experiments

Broad capability versus contextual adaptation. Figure 6 demonstrates that ECHO-PRE supports a broad repertoire of whole-body behaviors with a single pretrained policy, ranging from ground and terrain interactions to highly dynamic aerial motions. The robot seamlessly handles

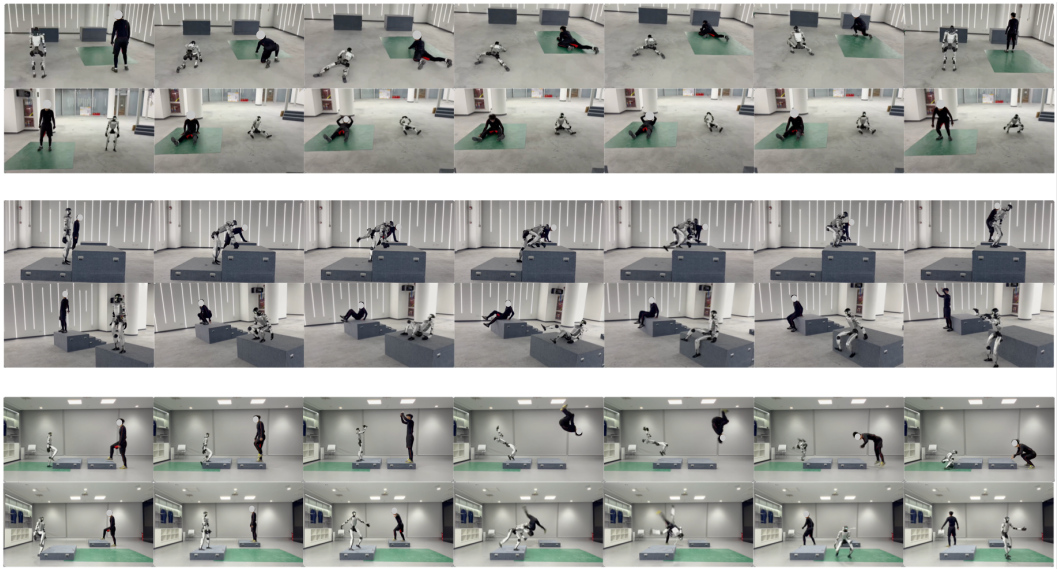

Figure 6: **Real-world whole-body teleoperation with ECHO-PRE.** A single pretrained policy enables the Unitree G1 to track human motions across ground interaction (top two rows), terrain interaction (middle two rows), and extreme acrobatics (bottom two rows). Examples include transitions between standing and floor-level postures, platform climbing and descent, and aerial flips. Each row shows successive frames of a teleoperated motion, ordered from left to right.



Figure 7: **Real-world in-context adaptation with ECHO. Top: manipulation.** The robot performs object manipulation with an added wrist payload; insets highlight the hand–object interactions. **Middle three rows: in-context whole-body behaviors.** Under payloads attached to its limbs, the robot executes diverse behaviors, including single-leg balance, squatting, sitting, and transitions between standing and prone postures. **Bottom: in-context adaptation process.** Successive frames show the robot's postural responses as a payload is attached to and removed from its arm. ECHO uses interaction-derived context to guide dynamics-specific control corrections without test-time parameter updates. Frames within each row progress from left to right.

Table 8: **Context-length scaling on LAFAN-40.** Bold and underlined values indicate the best and second-best results in each column, ranked before rounding. Arrows indicate the preferred direction. The 50/200-step runs rebuilt simulator/replay state at update 1,000.

Method	Linear Env. Acc. (%) ↑	Cross-motion mAP (%) ↑	Silhouette ↑	Payload mass R^2 ↑	Payload CoM R^2 ↑	Friction R^2 ↑
Short only (50)	91.967 ± 0.567	75.778 ± 2.225	0.2764 ± 0.0130	0.5037 ± 0.0054	0.0253 ± 0.0037	0.6899 ± 0.0169
Context 100	94.092 ± 0.342	84.132 ± 1.571	0.3321 ± 0.0131	0.6542 ± 0.0055	0.0751 ± 0.0043	0.8007 ± 0.0123
Context 200	96.657 ± 0.572	90.959 ± 1.318	0.4183 ± 0.0126	0.7842 ± 0.0094	0.0970 ± 0.0046	0.8834 ± 0.0059
Context 300	97.468 ± 0.632	93.583 ± 1.716	0.4484 ± 0.0140	0.8284 ± 0.0143	<u>0.1275</u> ± 0.0069	0.9146 ± 0.0102
Context 350	97.946 ± 0.791	94.776 ± 1.757	0.4693 ± 0.0136	0.8404 ± 0.0110	**0.1286** ± 0.0058	0.9269 ± 0.0087
Context 400	<u>98.135</u> ± 0.838	<u>95.276</u> ± 1.833	<u>0.4768</u> ± 0.0126	<u>0.8550</u> ± 0.0084	0.1263 ± 0.0065	<u>0.9331</u> ± 0.0086
Context 450	**98.403** ± 0.596	**96.165** ± 1.414	**0.4964** ± 0.0131	**0.8618** ± 0.0059	0.1209 ± 0.0058	**0.9352** ± 0.0071

Table 9: **Real-robot evaluation of whole-body motion tracking.** We report MPJPE-L (mm, ↓; mean ± sample standard deviation) on four representative motions, with each checkpoint evaluated over 10 trials per motion. All numerical entries correspond to 10/10 successful trials (100% SR), whereas *Fail* denotes 0/10 successful trials (0% SR). The two groups evaluate motion tracking without additional payload (top) and with a 2 kg payload attached to each lower leg (bottom). Bold and underlined values indicate the best and second-best mean tracking errors for each motion within each setting, respectively.

Method	Charleston	Single-leg balance	Sit down / stand up	Prone crawl / stand up
No additional payload				
HEFT	51.47 ± 3.68	42.53 ± 2.66	*Fail*	*Fail*
SONIC v1.1	54.81 ± 4.93	45.21 ± 3.61	*Fail*	65.84 ± 3.70
ECHO-PRE	<u>48.50</u> ± 3.21	<u>42.13</u> ± 2.45	<u>53.70</u> ± 4.79	<u>55.74</u> ± 4.39
ECHO-N	**47.98** ± 3.42	**39.69** ± 2.20	**49.73** ± 3.90	**53.65** ± 3.49
Added payload: 2 kg per lower leg				
Vanilla	64.23 ± 3.01	<u>43.52</u> ± 1.42	<u>51.51</u> ± 3.64	<u>69.52</u> ± 2.35
RMA-Student	<u>57.18</u> ± 3.56	45.39 ± 2.51	75.58 ± 3.04	77.74 ± 2.75
ECHO-H	**52.07** ± 2.21	**41.79** ± 2.13	**49.95** ± 3.55	**54.41** ± 2.57

substantial variations in posture, support conditions, and motion dynamics, including floor-level transitions, platform traversal, and aerial flips, highlighting the breadth and reusability of the learned behavioral capabilities.

Figure 7 further demonstrates that ECHO can adapt these capabilities to changing physical contexts through in-context interaction. Under additional wrist and limb payloads, the robot maintains precise manipulation and diverse whole-body behaviors, including balance, squatting, sitting, and posture transitions. Crucially, when the payload is attached or removed, ECHO adjusts its control response based on the resulting interaction history without updating model parameters at test time. This shows that ECHO complements broad pretrained capabilities with context-dependent dynamics adaptation, enabling precise behavior under previously unseen physical conditions.

A.6 LIMITATIONS

Despite its broad whole-body capabilities and effective in-context dynamics adaptation, ECHO has several limitations. First, context learning relies on simulated rollouts and privileged supervision, and generalization to physical effects beyond the training randomization requires further evaluation. Second, our memory mechanism assumes fixed physical parameters within each session and clears the history at session boundaries. Adaptation to unannounced changes within a session, where accumulated context may become outdated, remains insufficiently explored. Finally, our evaluation focuses on the Unitree G1 and reference-conditioned whole-body control; transfer across embodiments and integration with autonomous task planning remain untested. We leave broader physical

generalization, change-aware context memory, and extensions across embodiments and tasks to future work.

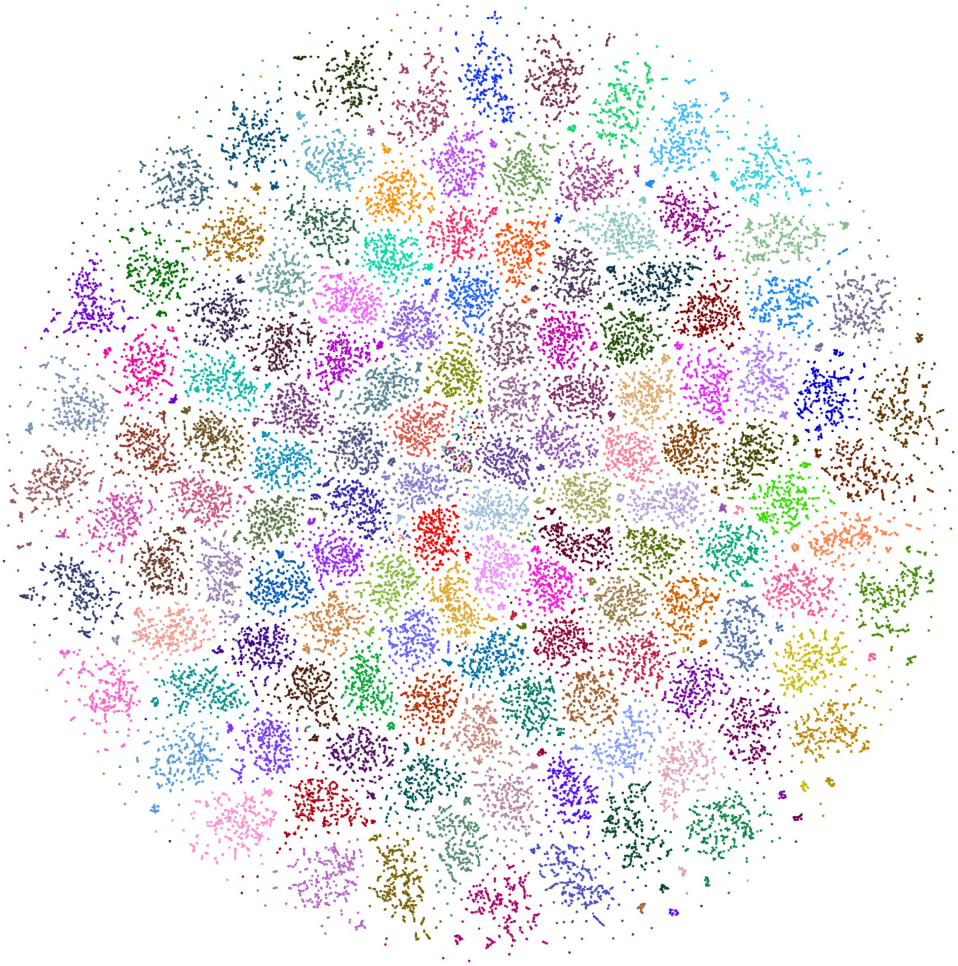

Figure 8: **Context representations on unseen AMASS motions.** t-SNE visualization of ECHO's context embeddings from 101,967 valid history windows collected across 100 AMASS motions and 128 physical environments. Each point represents an individual history-window embedding, without motion-level averaging, and each color denotes a fixed randomized physical configuration. The environment-wise grouping provides qualitative evidence that the learned representations distinguish physical conditions consistently across unseen motions.