Robust and Generalized Humanoid Motion Tracking

Yubiao Ma[1]†, Han Yu[2]†, Jiayin Xie[2], Changtai Lv[2], Qiang Luo[2], Chi Zhang[2], Yunpeng Yin[2], Boyang Xing[2], Xuemei Ren[1], and Dongdong Zheng[1,2]∗

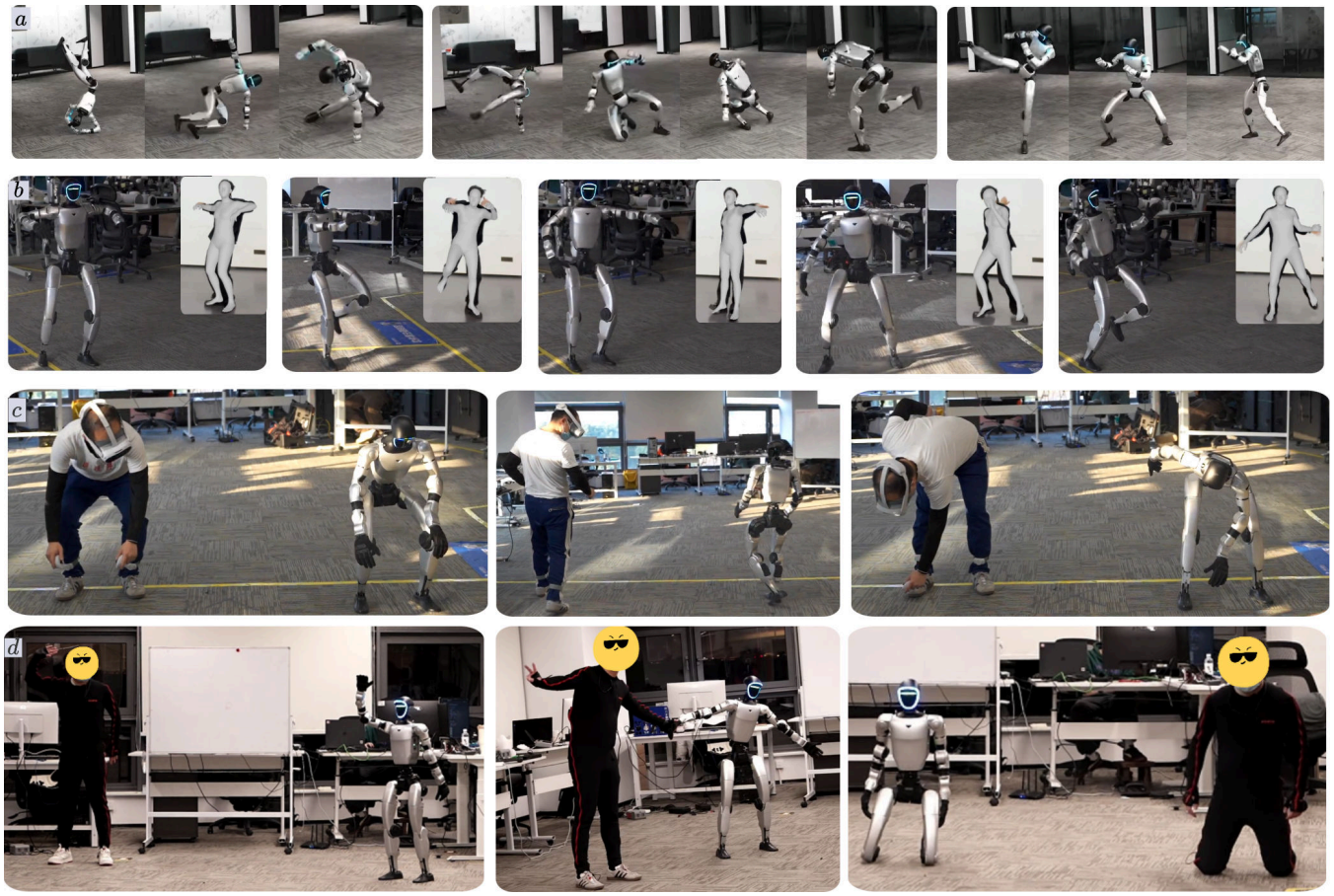

Fig. 1: Qualitative results illustrating the generalization of our method across different motion data sources, including public MoCap datasets, video-derived motions obtained from human pose estimation on public videos, and real-time teleoperation demonstrations via VR and a motion-capture suit.

***Abstract*— Learning a general humanoid whole-body controller is challenging because practical reference motions can exhibit noise and inconsistencies after being transferred to the robot domain, and local defects may be amplified by closed-loop execution, causing drift or failure in highly dynamic and contact-rich behaviors. We propose a dynamics-conditioned command aggregation framework that uses a causal temporal encoder to summarize recent proprioception and a multi-head cross-attention command encoder to selectively aggregate a context window based on the current dynamics. We further integrate a fall recovery curriculum with random unstable initialization and an annealed upward assistance force to improve robustness and disturbance rejection. The resulting policy requires only about 3.5 hours of motion data and supports single-stage end-to-end training without distillation. The proposed method is evaluated under diverse reference inputs and challenging motion regimes, demonstrating zero-shot transfer to unseen motions as well as robust sim-to-real transfer on a physical humanoid robot.**
Project page: `https://zeonsunlightyu.github.io/RGMT.github.io/`

[1]Beijing Institute of Technology, Beijing, China.
[2]Humanoid Robotics (Shanghai) Co., Ltd., Shanghai 201203, China.
†These authors contributed equally.
∗Corresponding author: Dongdong Zheng (`ddzheng@bit.edu.cn`).

I. Introduction

Humanoid robots are compelling largely because of their generality. Their morphology is naturally compatible with human environments, allowing them to operate within existing infrastructure and manipulate tools, workspaces, and interfaces designed for people. Moreover, their high-dimensional actuation and multi-contact capabilities support a broad spectrum of behaviors spanning locomotion, manipulation, and interaction [1]–[10]. To translate this vision into reliable operation, it is necessary to develop a general

whole-body controller that can coordinate the full body under changing contacts and task demands, while maintaining stable behavior over long horizons. However, this objective remains difficult to achieve in existing whole-body control research. High-fidelity imitation in practice still often relies on training for a single motion or a small set of motions [11]–[14], which tightly couples policy capability to a specific reference distribution and limits unified modeling and generalization across skills.

To learn diverse humanoid motions within a single policy, recent work has moved toward unified whole-body control that combines large-scale motion tracking objectives with broader motion coverage and more practical data acquisition. Several systems [15]–[18] focus on scalable human to humanoid data collection and teleoperation, providing richer and more diverse demonstrations for training. In parallel, a number of motion tracking based controllers [19]–[22] aim to directly train a universal tracker that can follow many motions under different disturbances. These approaches demonstrate promising progress toward general whole-body controllers on humanoid robots, but their tracking accuracy and closed loop stability remain suboptimal, especially during highly dynamic maneuvers and rapid contact transitions. SONIC [23] further pushes motion coverage and naturalness, but it relies on very large-scale data and training resources, using more than 700 hours of motion data and substantial compute, which raises the barrier for iterative research and deployment. Moreover, long-horizon operation in the real world requires not only stable tracking but also recovery after disturbances or falls, yet fall recovery is often not integrated into the main control policy [6], [7], limiting robustness and safety. Therefore, we need a general whole-body control framework that can learn effectively under limited data, maintain stable closed-loop tracking under imperfect references, and jointly integrate robustness improvements for highly dynamic and multi-contact scenarios with fall recovery within a single training and policy pipeline, enabling safe and continuous real-world task execution.

This paper presents a learning framework for general humanoid whole-body control. The key idea is to condition the policy on the current dynamics, enabling it to interpret and aggregate contextual commands selectively rather than treating all reference signals as equally reliable supervision. We obtain a compact dynamics representation from recent proprioceptive history using a causal temporal encoder and use it to guide command aggregation via multi-head attention [24]. This design allows the policy to adaptively select and adjust reference segments under feasibility constraints imposed by the current dynamics, reducing the impact of noise and artifacts, particularly for highly dynamic motions and frequent contact transitions. To further support safe and continuous real-world operation, we incorporate fall recovery into the same training framework [6], [7]. This integration broadens the experienced state distribution, improves robustness to disturbances, and strengthens tracking performance for contact-rich motion segments. Our contributions are summarized as follows:

- We propose a dynamics-conditioned command aggregation framework that combines causal temporal dynamics encoding with multi-head cross-attention. This design enables selective use of contextual commands under imperfect references and improves tracking accuracy and closed-loop stability in highly dynamic and contact-rich scenarios. The resulting general policy is trained end-to-end using a compact dataset of about 3.5 hours, without distillation or multi-stage training.
- We integrate fall recovery into a unified training framework. With randomized unstable initialization and an annealed external assistance force, a single policy learns stable control and self-recovery over a broader state distribution and contact conditions, leading to significantly improved robustness and disturbance rejection.
- We demonstrate strong generalization across diverse reference sources, including mocap, video-derived motions, and real-time full-body teleoperation. The learned policy transfers zero-shot to unseen motions and deploys robustly on the Unitree G1, enabling stable long-horizon tracking with integrated recovery and downstream applications such as joystick-driven locomotion.

II. METHOD

A. Humanoid Motion Dataset

Our motion corpus is constructed from LAFAN1 [25] and a selected subset of AMASS [26], and all sequences are retargeted to our humanoid using General Motion Retargeting [27]. In practice, large-scale mocap sources [26] and their retargeted counterparts often exhibit substantial redundancy and may include segments with low motion quality and inconsistent contacts. We therefore perform quality control to remove infeasible motions and low-quality sequences and to reduce redundancy, resulting in a relatively compact dataset of approximately 3.5 hours.

This quality-driven construction is crucial for general humanoid motion tracking. Even when the raw corpus is large, effective supervision can be limited by duplicated motions and low-quality clips, which can distract optimization and reduce training efficiency. In contrast, a smaller but diverse and higher-quality reference set provides cleaner and more informative supervision, improving generalization and closed-loop tracking accuracy. Importantly, enabled by our dynamics-conditioned command aggregation, this compact dataset is sufficient to train a strong general whole-body policy that is robust to noisy references and generalizes effectively to unseen motions.

B. Motion Tracking Formulation

1) Observation Space: Our policy receives an observation $\mathbf{o}_t$ that consists of a proprioceptive component and a command component. The proprioceptive observation is

$$\mathbf{o}_t = \Big[\mathbf{g}_t^{\mathrm{proj}},\ \boldsymbol{\omega}_t,\ \mathbf{q}_t - \mathbf{q}_0,\ \dot{\mathbf{q}}_t,\ \mathbf{a}_{t-1}\Big], \tag{1}$$

where $\mathbf{g}_t^{\mathrm{proj}} \in \mathbb{R}^3$ denotes the gravity direction projected into the body frame, $\boldsymbol{\omega}_t \in \mathbb{R}^3$ is the base angular velocity,

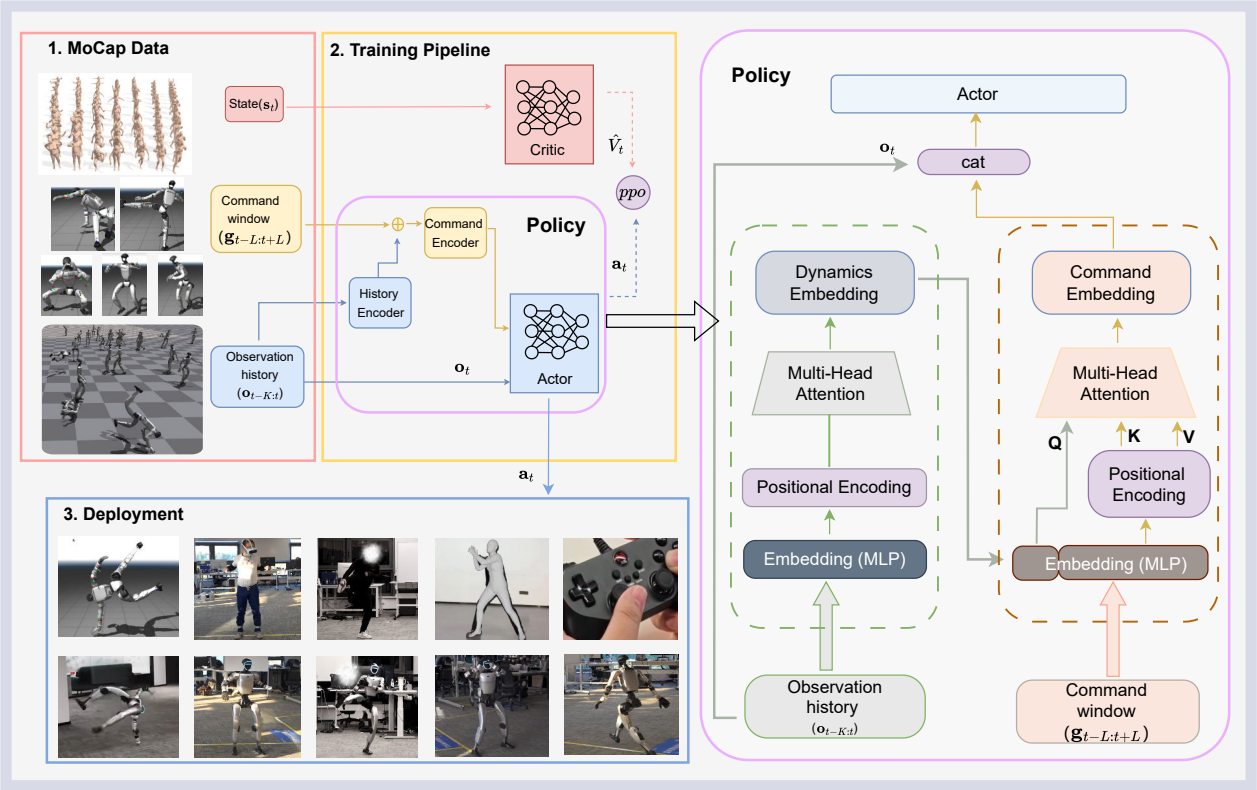

Fig. 2: Overview of the proposed whole-body control pipeline. A history encoder extracts a dynamics embedding from recent proprioception, which conditions a command encoder to aggregate the contextual command window. The resulting representation is fused with the current observation and fed to an actor-critic policy trained with PPO, and the learned actor is deployed for real-world whole-body motion tracking and teleoperation.

$\mathbf{q}_t \in \mathbb{R}^{29}$ and $\dot{\mathbf{q}}_t \in \mathbb{R}^{29}$ are joint positions and velocities, $\mathbf{q}_0$ is the default joint configuration, and $\mathbf{a}_{t-1} \in \mathbb{R}^{29}$ is the previous action.

The command observation provides per step reference targets extracted from the reference motion,

$$\mathbf{g}_t = \left[\mathbf{v}_t^{\text{ref}},\ \boldsymbol{\omega}_t^{\text{ref}},\ \mathbf{g}_t^{\text{ref}},\ \mathbf{q}_t^{\text{ref}}\right], \tag{2}$$

where $\mathbf{v}_t^{\text{ref}} \in \mathbb{R}^3$ and $\boldsymbol{\omega}_t^{\text{ref}} \in \mathbb{R}^3$ are reference base linear and angular velocities expressed in the body frame, $\mathbf{g}_t^{\text{ref}} \in \mathbb{R}^3$ is the reference gravity direction in the body frame, and $\mathbf{q}_t^{\text{ref}} \in \mathbb{R}^{29}$ is the reference joint position at time t.

For value learning, we adopt an asymmetric actor-critic training scheme. The critic additionally takes a privileged observation that facilitates more accurate value estimation,

$$\mathbf{o}_t^{\text{priv}} = \left[h_t^{\text{ref}},\ \mathbf{x}_t^{\text{link}},\ \mathbf{v}_t\right], \tag{3}$$

where h_t^{ref} denotes the reference base height, $\mathbf{x}_t^{\text{link}}$ denotes body link poses, and $\mathbf{v}_t$ is the base linear velocity. Accordingly, the critic input is

$$\mathbf{s}_t = \left[\mathbf{o}_t,\ \mathbf{g}_t,\ \mathbf{o}_t^{\text{priv}}\right]. \tag{4}$$

All components of $\mathbf{s}_t$ used by the critic are noise-free, while the actor receives noisy observations $\mathbf{o}_t$.

2) Action Space and Low-level Control: The policy outputs a residual joint position command $\mathbf{a}_t \in \mathbb{R}^{29}$. We interpret $\mathbf{a}_t$ as a corrective offset to the reference joint configuration and form the PD setpoint as

$$\mathbf{q}_t^{\text{tar}} = \mathbf{q}_t^{\text{ref}} + \mathbf{a}_t. \tag{5}$$

Joint torques are computed by a joint-level PD controller,

$$\boldsymbol{\tau}_t = \mathbf{K}_p\left(\mathbf{q}_t^{\text{tar}} - \mathbf{q}_t\right) - \mathbf{K}_d\,\dot{\mathbf{q}}_t, \tag{6}$$

where $\mathbf{K}_p$ and $\mathbf{K}_d$ are diagonal gain matrices.

This residual formulation improves tracking accuracy by anchoring the PD setpoint to the reference motion while allowing corrective adjustments. It also makes exploration more efficient, since the policy searches around the reference pose rather than over the full joint configuration space. As a result, training becomes more sample efficient and converges faster in practice.

3) Reward Function: Following [14], we adopt a dense reward that combines reference tracking with safety and smoothness regularization. We define keypoints as a fixed set of links used in [14], and formulate the tracking reward on these keypoints. Specifically, the tracking part measures keypoint alignment, relative pose consistency, and keypoint

velocity consistency with exponential kernels. The regularization part penalizes rapid action changes, joint limit violations, and contacts on non-target body parts, which stabilizes learning and reduces physically implausible behaviors.

C. Policy Learning Framework

Multi-head attention (MHA) [24] provides an effective mechanism for query-conditioned information fusion. It computes content-based similarities between a query and a set of candidate features to produce adaptive aggregation weights, yielding a compact representation that emphasizes relevant elements while attenuating irrelevant or noisy ones. The multi-head formulation performs this matching in multiple learned subspaces, allowing the model to capture diverse relevance cues in parallel and increasing expressiveness beyond a single attention map. This capability is well suited to whole-body control, where command inputs are high dimensional and heterogeneous and may contain unreliable segments.

In practice, command sequences obtained from diverse sources often contain artifacts such as body penetration, inconsistent contacts, and high frequency noise. If the policy treats the entire command sequence as equally trustworthy supervision, abnormal segments can enter the representation with the same weight and be amplified in the action output, which can ultimately harm closed loop stability. Motivated by this observation, we introduce dynamics-conditioned cross-attention in the policy architecture, as illustrated in Fig. 2. A dynamics representation is first extracted from recent proprioceptive history using causal temporal encoding and used as the query signal, and the contextual command window is then aggregated with adaptive attention weights. This design enables the policy to interpret and filter command information under physical feasibility constraints, reducing the influence of inconsistent reference signals on the control representation.

1) History Encoder: We encode the most recent 10-step proprioceptive observations into a compact dynamics embedding. The input sequence is

$$\mathbf{o}_{t-K:t} = \big[\mathbf{o}_{t-K}, \dots, \mathbf{o}_t\big], \tag{7}$$

where each $\mathbf{o}_t \in \mathbb{R}^{93}$ and the sequence length is $K+1$.

Each observation is mapped to a token in an embedding space of dimension $n_{\text{embd}} = 128$ by a two-layer multi-layer perceptron (MLP),

$$\mathbf{E}_{t-K:t} = \text{MLP}(\mathbf{o}_{t-K:t}) \in \mathbb{R}^{(K+1)\times n_{\text{embd}}}, \tag{8}$$

and a sinusoidal positional encoding is added to preserve temporal order,

$$\tilde{\mathbf{E}}_{t-K:t} = \mathbf{E}_{t-K:t} + \mathbf{P}, \qquad \mathbf{P} \in \mathbb{R}^{(K+1)\times n_{\text{embd}}}. \tag{9}$$

The resulting token sequence is processed by a lightweight causal Transformer with multi-head self-attention. Causality is enforced by a causal mask so that the token at time τ can attend only to tokens from times $\leq \tau$ within the window. Letting $\mathbf{H}^{(0)} = \tilde{\mathbf{E}}_{t-K:t}$, the causal self-attention block yields

$$\begin{aligned} \mathbf{H}^{(1)} &= \mathbf{H}^{(0)} + \text{MHA}\Big(\text{LN}\Big(\mathbf{H}^{(0)}\Big)\Big), \\ \mathbf{H}^{(2)} &= \mathbf{H}^{(1)} + \text{MLP}\Big(\text{LN}\Big(\mathbf{H}^{(1)}\Big)\Big), \\ \bar{\mathbf{H}} &= \text{LN}\Big(\mathbf{H}^{(2)}\Big), \end{aligned} \tag{10}$$

where $\text{LN}(\cdot)$ denotes layer normalization. Finally, we aggregate token features over time via element-wise max pooling to obtain the dynamics embedding $\mathbf{h}_t \in \mathbb{R}^{n_{\text{embd}}}$:

$$\mathbf{h}_t[j] = \max_{\tau \in \{t-K,\dots,t\}} \bar{\mathbf{H}}_\tau[j], \qquad j = 1, \dots, n_{\text{embd}}. \tag{11}$$

This embedding extracts the robot dynamics from recent proprioceptive history and is used as the query signal for the subsequent dynamics-conditioned command encoder.

2) Command Encoder: The command encoder compresses a contextual command window into a compact latent representation while conditioning the aggregation on the current dynamics. Its inputs are the dynamics embedding $\mathbf{h}_t \in \mathbb{R}^{n_{\text{embd}}}$ and the command sequence

$$\mathbf{g}_{t-L:t+L} = \big[\mathbf{g}_{t-L}, \dots, \mathbf{g}_{t+L}\big], \tag{12}$$

where each $\mathbf{g}_t \in \mathbb{R}^{38}$ and the window length is $2L+1$. The dynamics embedding is projected to the Transformer dimension through a two-layer multi-layer perceptron, yielding the query vector

$$\mathbf{q}_t = \text{MLP}_{\text{dyn}}(\mathbf{h}_t) \in \mathbb{R}^{n_{\text{embd}}}. \tag{13}$$

In parallel, the command window is mapped to a token sequence in the same feature space using another two-layer multi-layer perceptron, and a sinusoidal positional encoding is added to preserve temporal order

$$\tilde{\mathbf{Z}} = \text{MLP}_{\text{cmd}}(\mathbf{g}_{t-L:t+L}) + \mathbf{P}^{\text{cmd}}, \qquad \tilde{\mathbf{Z}} \in \mathbb{R}^{(2L+1)\times n_{\text{embd}}}. \tag{14}$$

The encoder then applies a single dynamics-conditioned cross-attention block to aggregate the command tokens into a compact latent representation

$$\begin{aligned} \mathbf{s}^{(1)} &= \mathbf{q}_t + \text{MHA}\Big(\text{LN}(\mathbf{q}_t)\,,\, \tilde{\mathbf{Z}}\Big), \\ \mathbf{s}^{(2)} &= \mathbf{s}^{(1)} + \text{MLP}\Big(\text{LN}\Big(\mathbf{s}^{(1)}\Big)\Big), \\ \mathbf{u}_t &= \text{LN}\Big(\mathbf{s}^{(2)}\Big) \in \mathbb{R}^{n_{\text{embd}}}. \end{aligned} \tag{15}$$

The resulting vector $\mathbf{u}_t$ serves as a compact, dynamics-conditioned command embedding at time t. Because the cross-attention weights are conditioned on $\mathbf{h}_t$ through the query $\mathbf{q}_t$, the encoder adaptively emphasizes command elements that are more consistent with the current dynamics and down-weights unreliable segments, thereby reducing the influence of abnormal command artifacts on the control representation.

TABLE I: Performance under different motion data sources. We report mean $\pm$ standard deviation. Higher is better for success rate, and lower is better for E_{MPJPE}.

Method	MoCap Data		Video-derived Motion		Ground-interaction Motion	
	Succ. $\uparrow$	$\mathbf{E}_{\mathrm{MPJPE}}$ $\downarrow$	**Succ.** $\uparrow$	$\mathbf{E}_{\mathrm{MPJPE}}$ $\downarrow$	**Succ.** $\uparrow$	$\mathbf{E}_{\mathrm{MPJPE}}$ $\downarrow$
(a) Baseline						
GMT	84.6%	65.15 ± 1.12	72.4%	96.47 ± 1.98	48.9%	146.95 ± 5.12
Any2Track	89.2%	56.96 ± 0.91	54.3%	112.16 ± 3.96	41.2%	209.57 ± 4.22
Ours	**98.3%**	$\mathbf{41.12 \pm 0.12}$	**94.6%**	$\mathbf{46.56 \pm 0.28}$	**90.1%**	$\mathbf{54.92 \pm 0.93}$
(b) Ablations on Policy Architecture						
Ours SelfAttn CmdEnc	89.8%	51.96 ± 0.72	76.7%	67.19 ± 1.31	73.2%	92.65 ± 2.31
Ours CNN HistEnc	94.3%	48.61 ± 0.32	91.9%	53.13 ± 0.78	81.5%	68.92 ± 1.84
Ours	**98.3%**	$\mathbf{41.12 \pm 0.12}$	**94.6%**	$\mathbf{46.56 \pm 0.28}$	**90.1%**	$\mathbf{54.92 \pm 0.93}$
(c) Ablations on Fall Recovery						
Ours w/o Fall Recovery	**98.4%**	$\mathbf{40.98 \pm 0.09}$	**94.9%**	$\mathbf{46.16 \pm 0.31}$	70.5%	96.75 ± 2.77
Ours	98.3%	41.12 ± 0.12	94.6%	46.56 ± 0.28	**90.1%**	$\mathbf{54.92 \pm 0.93}$

D. Fall Recovery Integration

Automatic fall recovery is a key prerequisite for reliable humanoid deployment. Without an effective self-recovery mechanism, the system not only faces significant safety risks but also requires frequent human intervention, breaking task continuity. This issue is particularly pronounced for whole-body motion tracking, where rapid dynamics and frequent contact transitions can amplify closed-loop errors and trigger instability. Therefore, we integrate a simple yet reliable automatic fall recovery mechanism into our whole-body control framework.

1) Randomized Recovery Initialization: We designate a subset of the parallel environments as recovery environments with probability 0.15 and reset the robot in these environments to randomized poses, exposing the policy to a broad range of unstable configurations and contact initial conditions. This process also enriches contact experience during training, since repeated falls and stand-ups naturally induce diverse ground-contact patterns, which improves tracking accuracy and generalization for motions with frequent or complex contact transitions.

For these recovery environments, we apply an upward pulling force with magnitude uniformly sampled from $[0, 200]$ to assist exploration at early training stages by increasing the frequency of recoverable states. The assistance is linearly annealed over training iterations and reduced to a negligible level, ensuring that the final policy performs fall recovery using only its own control.

2) Termination Conditions: We employ state-based episode termination and environment resets to maintain stable training and high-quality rollouts. For all environments, an episode is reset either when it reaches a predefined time limit or when an instability event is detected. Instability is identified by three conditions: excessive base orientation deviation, insufficient base height, and abnormally low height of key body links.

To learn automatic fall recovery, we introduce an additional termination strategy for the recovery environments. Within a predetermined recovery window of 3 seconds, recovery environments are not terminated early by the instability criteria, allowing the policy to complete stand-up and re-stabilization within the same episode. If the robot fails to recover within this window, the episode is terminated and the environment is reset, avoiding prolonged rollouts in unrecoverable states and maintaining training efficiency.

TABLE II: Noise specifications for command.

Command	Noise Specification
Base linear velocity jitter	$\Delta \mathbf{v}_t^{\mathrm{ref}} \sim \mathcal{U}([-0.5, 0.5]^3)$ (z: ± 0.2)
Base angular velocity jitter	$\Delta \omega_t^{\mathrm{ref}} \sim \mathcal{U}([-0.52, 0.52]^3)$
Base gravity direction jitter	$\Delta \mathbf{g}_t^{\mathrm{ref}} \sim \mathcal{U}([-0.05, 0.05]^3)$
Base joint position jitter	$\Delta \mathbf{q}_t^{\mathrm{ref}} \sim \mathcal{U}([-0.1, 0.1])$

E. Training Setup

We train our policy in the Isaac Gym simulator [28]. All training runs are conducted on a single NVIDIA RTX 4090 GPU with 5,680 parallel environments. We train on approximately 3.5 hours of motion data curated from subsets of LAFAN1 [25] and AMASS [26], and observe strong generalization to previously unseen motions. The proprioceptive history length is set to $K = 9$, and the command window half-length is set to $L = 10$.

III. Experiments

In this section, we evaluate our approach on the 29 degrees-of-freedom (DOF) Unitree G1 [29] humanoid robot and demonstrate strong generalization and robustness in both simulation and the real world. We conduct quantitative comparisons against representative baselines and targeted ablation studies to validate improved whole-body motion tracking performance and robustness to noise in command inputs. Finally, we deploy the learned policy on the physical robot to showcase reliable tracking and generalization across diverse motions, and demonstrate its applicability to downstream tasks such as real-time teleoperation and online motion generation.

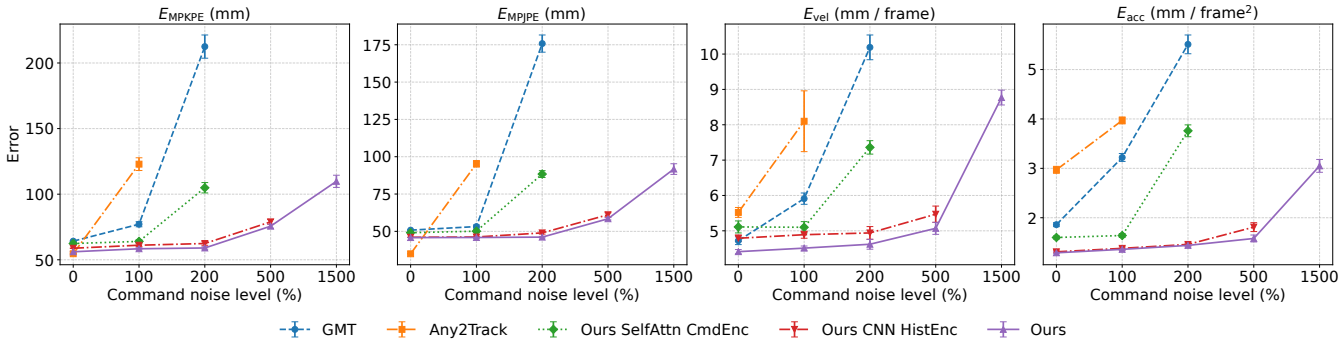

Fig. 3: Robustness under reference command noise.

A. Performance Evaluation

1) Comparison with Baselines: We validate that the proposed approach learns a general motion tracker that achieves stable and accurate tracking over a wide range of previously unseen motions. We compare our method against two representative recent trackers, Any2Track [20] and GMT [19], both of which provide officially released models. To ensure consistency and fairness, we directly use these models and evaluate all methods on the same datasets. All methods are tested in MuJoCo [30], which is supported by the above baseline implementations and provides a unified evaluation platform. A comprehensive set of metrics is adopted to capture both pose accuracy and physical feasibility. Specifically, the success rate (Succ) is defined as the fraction of rollouts without falls, where a fall is identified when the root height deviates from the reference by more than 0.2 m. In addition, we report the mean per-joint position error (MPJPE) E_{MPJPE} (in mm), which measures the 3D position error of joints and quantifies joint-level tracking accuracy.

Table I (a) compares baseline performance across three motion sources: MoCap data curated from subsets of LAFAN1 and AMASS, video-derived motions, and ground-interaction motions. The MoCap subset comes from the same MoCap sources used to train all methods and thus measures in-distribution tracking performance. The video-derived motions are estimated from publicly available videos and cover common behaviors such as walking, dance, and martial arts. Since this subset is not used for training, it directly evaluates cross-source generalization. Across all three subsets, our method achieves the highest success rate and the lowest E_{MPJPE}. The improvement is most pronounced on the video-derived subset, where we substantially outperform GMT and Any2Track, indicating stronger robustness to distribution shifts in the reference motions.

2) Ablations on Policy Architecture: Table I (b) studies the effect of key design choices in our tracker. Replacing the causal history encoder with a CNN-based variant (*Ours CNN HistEnc*) consistently degrades performance across all three data sources, indicating that the causal history encoder contributes to stable and accurate tracking. Replacing the cross-attention command encoder with a self-attention variant (*Ours SelfAttn CmdEnc*) leads to a much larger performance drop, particularly on the video-derived and ground-interaction subsets. This highlights the key role of dynamics-conditioned cross-attention in robust command aggregation under distribution shifts and command artifacts.

3) Ablations on Fall Recovery: Table I (c) studies the effect of integrating fall recovery during training. Incorporating fall recovery does not noticeably compromise tracking accuracy or generalization to unseen motions, as reflected by comparable performance on the MoCap and video-derived subsets. In contrast, removing fall recovery training (*Ours w/o Fall Recovery*) substantially degrades performance on the ground-interaction subset, which includes contact-intensive behaviors such as crawling, kneeling, sitting, and breakdance-style motions. In this setting, the success rate (Succ.) drops markedly and E_{MPJPE} increases. These results suggest that fall recovery not only provides self-recovery capability for safer long-horizon execution but also broadens contact experience during training, thereby improving robustness under complex ground interactions.

B. Robustness Evaluation

Robustness to command noise is essential for real-time teleoperation and generalization to unseen motions. In practical applications, commands from human operators, motion estimation pipelines, or high-level planners are often noisy and subject to uncertainty. To evaluate robustness under such conditions, we evaluate on the Charleston dance motion, which is included in the training set of all methods and has been showcased by both GMT and Any2Track. We inject varying levels of noise into the reference commands to systematically assess the robustness of each method. The noise specifications (base noise patterns) are summarized in Table II . We define the noise level (%) by uniformly scaling the ranges in Table II .

In Fig. 3, we report additional metrics beyond E_{MPJPE}. E_{MPKPE} measures the mean keypoint position error (in mm). To assess physical fidelity, E_{vel} and E_{acc} measure the differences in keypoint velocities and accelerations relative to the reference motion, reported in mm/frame and mm/frame^2, respectively. As shown in Fig. 3, baseline methods such as GMT and Any2Track degrade rapidly as the noise level increases. E_{MPJPE}, E_{vel}, and E_{acc} rise substantially, and the baselines struggle to maintain stable tracking when the

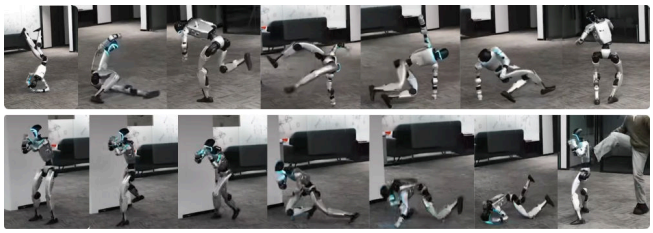

Fig. 4: Real-world dance tracking with fall recovery.

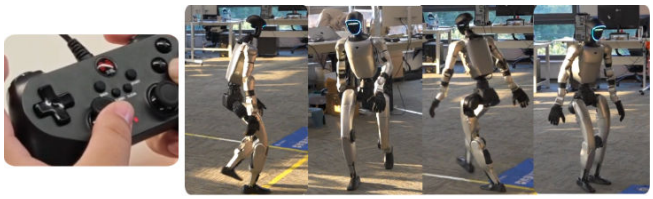

Fig. 5: Joystick-driven stylized locomotion.

noise level exceeds 200%. These results suggest that, without an explicit mechanism to interpret and filter corrupted commands, baseline methods are more likely to amplify command noise in closed-loop execution, leading to instability at high noise levels.

In contrast, our method demonstrates significantly stronger robustness across all four metrics. Even at high noise levels, the errors remain relatively low and degrade much more gradually. Ablation experiments further confirm this trend. Replacing the causal history encoder with a CNN-based variant consistently degrades robustness, confirming the importance of the causal history encoder. Replacing the cross-attention command encoder with a self-attention variant leads to a much larger degradation, especially in the high-noise regime, highlighting the cross-attention module as the key mechanism for filtering noisy commands. Notably, our full model maintains stable motion tracking even under noise levels up to 1500%, demonstrating strong tolerance to severe command corruption. Overall, these results highlight the central role of dynamics-conditioned cross-attention in filtering noisy commands and improving robustness, while the causal history encoder provides complementary benefits by stabilizing the dynamics representation used for command interpretation.

C. Real-World Experiments

1) Robust Whole-Body Motion Tracking: Fig. 4 qualitatively illustrates two challenging behaviors enabled by our policy. In the top row, the robot tracks a breakdance-style motion with frequent ground contacts and rapid contact transitions, demonstrating our policy's ability to robustly coordinate whole-body motion under complex contact patterns. In the bottom row, we apply an external push that causes the robot to fall. The policy autonomously executes a recovery maneuver and then resumes the motion-tracking task without manual resets, which improves robustness for long-horizon deployment and reduces the need for human intervention.

2) Video-Derived Motion Tracking: Fig. 1(b) presents qualitative results on video-derived motions to evaluate generalization to unseen motions. For these examples, the reference commands are produced by a video-based human motion estimation pipeline [31], which introduces noise and distribution shift. Despite these challenges, our policy tracks the commanded motions on the real robot with high fidelity, reproducing the overall timing and whole-body coordination. These results demonstrate effective transfer to previously unseen motion styles.

3) Real-Time Teleoperation: We develop a real-time full-body teleoperation pipeline with two different motion-tracking front ends: a consumer-grade PICO VR whole-body tracking interface and a wearable motion-capture suit. For the PICO setup, the operator wears a VR headset, two ankle trackers, and handheld controllers, and the system outputs full-body pose estimates at runtime. For the motion-capture suit setup, the suit streams full-body pose measurements online through its vendor SDK, which are converted to the same kinematic representation to ensure a unified downstream interface. In both cases, the estimated human motion is streamed in real-time and transformed into our reference command representation $\mathbf{g}_t$, which is then consumed by the deployed actor together with proprioceptive observations. This teleoperation setting is substantially noisier than offline MoCap due to sensing drift, latency, and operator inconsistency, and thus serves as a practical stress test of generalization and robustness. As shown in Fig. 1(c) and (d), under both teleoperation sources, our policy remains stable and precisely tracks the incoming commands without manual resets, enabling challenging whole-body behaviors such as crawling, high kicks, and deep squats.

4) Joystick-driven locomotion: Fig. 5 illustrates a downstream integration of our tracker into a representative computer graphics motion synthesis method [32]. A handheld game joystick provides high-level locomotion commands, which are mapped to a sequence of reference motion targets and streamed to our policy as contextual commands. Although the synthesized reference sequence can exhibit abrupt transitions due to discrete clip-switching and matching artifacts, our policy remains stable and tracks the commanded base velocity in a stylized manner, producing coherent whole-body locomotion on the real robot. This result suggests that our tracker can serve as a robust low-level controller for upstream motion generation modules, tolerating non-smooth reference trajectories while preserving responsive velocity tracking and motion style.

IV. Conclusion

We propose a whole-body motion tracking framework that achieves robust, generalized humanoid control with a single policy. By conditioning command aggregation on a dynamics representation extracted from recent proprioception, the policy can down-weight inconsistent reference segments and remain stable under substantial command corruption. Results in simulation and on a 29-DoF Unitree G1 demonstrate accurate tracking across diverse motion sources, strong generalization to previously unseen motions,

and reliable execution under complex contact patterns and external disturbances, enabling practical applications such as real-time teleoperation and joystick-driven locomotion.

Future work will incorporate global localization to enable long-horizon, world-frame consistent tracking, and expand integration with upstream motion generation and planning modules to support richer downstream tasks.

REFERENCES

[1] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik, "Humanoid locomotion as next token prediction," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 79 307–79 324.

[2] I. Radosavovic, S. Kamat, T. Darrell, and J. Malik, "Learning humanoid locomotion over challenging terrain," 2024.

[3] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "Beamdojo: Learning agile humanoid locomotion on sparse footholds," in *Proc. Robotics Sci. Syst. (RSS)*, Los Angeles, CA, USA, Jun. 2025.

[4] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through reinforcement learning," in *Proc. Robotics Sci. Syst. (RSS)*, Daegu, Republic of Korea, Jul. 2023.

[5] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," in *Proc. Conf. Robot Learn. (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, Nov. 2025, pp. 1975–1991.

[6] T. Huang, J. Ren, H. Wang, Z. Wang, Q. Ben, M. Wen, X. Chen, J. Li, and J. Pang, "Learning humanoid standing-up control across diverse postures," in *Proc. Robotics Sci. Syst. (RSS)*, Los Angeles, CA, USA, Jun. 2025.

[7] X. He, R. Dong, Z. Chen, and S. Gupta, "Learning getting-up policies for real-world humanoid robots," in *Proc. Robotics Sci. Syst. (RSS)*, Los Angeles, CA, USA, Jun. 2025.

[8] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Proc. Robotics Sci. Syst. (RSS)*, Delft, Netherlands, Jul. 2024.

[9] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," in *Proc. Robotics Sci. Syst. (RSS)*, Los Angeles, CA, USA, Jun. 2025.

[10] Y. Zhang, Y. Yuan, P. Gurunath, T. He, S. Omidshafiei, A. Aghamohammadi, M. Vazquez-Chanlatte, L. Pedersen, and G. Shi, "Falcon: Learning force-adaptive humanoid loco-manipulation," 2025.

[11] T. Zhang, B. Zheng, R. Nai, Y. Hu, Y. J. Wang, G. Chen, F. Lin, J. Li, C. Hong, K. Sreenath, and Y. Gao, "Hub: Learning extreme humanoid balance," in *Proc. Conf. Robot Learn. (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, Sep. 2025, pp. 686–704.

[12] W. Xie, J. Han, J. Zheng, H. Li, X. Liu, J. Shi, W. Zhang, C. Bai, and X. Li, "Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

[13] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbabu, C. Pan, Z. Yi, G. Qu, K. Kitani, J. K. Hodgins, L. Fan, Y. Zhu, C. Liu, and G. Shi, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," in *Proc. Robotics Sci. Syst. (RSS)*, Los Angeles, CA, USA, Jun. 2025.

[14] T. E. Truong, Q. Liao, X. Huang, G. Tevet, C. K. Liu, and K. Sreenath, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," 2025.

[15] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Proc. Conf. Robot Learn. (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, Nov. 2025, pp. 1516–1540.

[16] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang, "Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks," in *Proc. Conf. Robot Learn. (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, Sep. 2025, pp. 4493–4505.

[17] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "Twist2: Scalable, portable, and holistic humanoid data collection system," 2025.

[18] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," in *RSS 2025 Workshop Whole-Body Control and Bimanual Manipulation*, Jun. 2025.

[19] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "Gmt: General motion tracking for humanoid whole-body control," 2025.

[20] Z. Zhang, J. Guo, C. Chen, J. Wang, C. Lin, Y. Lian, H. Xue, Z. Wang, M. Liu, J. Lyu, H. Liu, H. Wang, and L. Yi, "Track any motions under any disturbances," 2025.

[21] K. Yin, W. Zeng, K. Fan, M. Dai, Z. Wang, Q. Zhang, Z. Tian, J. Wang, J. Pang, and W. Zhang, "Unitracker: Learning universal whole-body motion tracker for humanoid robots," 2025.

[22] J. Han, W. Xie, J. Zheng, J. Shi, W. Zhang, T. Xiao, and C. Bai, "Kungfubot2: Learning versatile motion skills for humanoid whole-body control," 2025.

[23] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z. Cao, J. Li, D. Minor, Q. Ben, X. Da, R. Ding, C. Hogg, L. Song, E. Lim, E. Jeong, T. He, H. Xue, W. Xiao, Z. Wang, S. Yuen, J. Kautz, Y. Chang, U. Iqbal, L. Fan, and Y. Zhu, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," 2025.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[25] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. J. Pal, "Robust motion in-betweening," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 60:1–60:12, 2020.

[26] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proc. IEEE CVF Int. Conf. Comput. Vis. (ICCV)*. IEEE, 2019, pp. 5441–5450.

[27] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," 2025.

[28] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu based physics simulation for robot learning," in *Proc. NeurIPS Track Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., 2021.

[29] Unitree Robotics, "Unitree g1: Humanoid robot functions and price," 2024, accessed 2025-10-31.

[30] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *Proc. IEEE RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Vilamoura-Algarve, Portugal: IEEE, Oct. 2012, pp. 5026–5033.

[31] Z. Shen, H. Pi, Y. Xia, Z. Cen, S. Peng, Z. Hu, H. Bao, R. Hu, and X. Zhou, "World-grounded human motion recovery via gravity-view coordinates," in *SIGGRAPH Asia 2024 Conference Papers (SA '24)*. Tokyo, Japan: ACM, Dec. 2024, pp. 144:1–144:11.

[32] D. Holden, O. Kanoun, M. Perepichka, and T. Popa, "Learned motion matching," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 53:1–53:12, 2020.