COMMANDING HUMANOID BY FREE-FORM LANGUAGE: A LARGE LANGUAGE-ACTION MODEL WITH UNIFIED MOTION VOCABULARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Enabling humanoid robots to follow free-form natural language commands is a critical step toward seamless human-robot interaction and general-purpose embodied AI. However, existing methods remain limited, often constrained to simple instructions or forced to sacrifice motion diversity for physical plausibility. To address this gap, we present Humanoid-LLA, a Large Language-Action model that translates abstract and complex instructions to executable whole-body humanoid behaviors. Our approach tackles two core challenges: paired language-humanoid motion data scarcity and physical instability. First, we bridge high-level language semantics with physically grounded control by learning a unified human-humanoid motion vocabulary. Second, we introduce a novel two-stage fine-tuning framework that begins with supervised motion CoT learning, followed by reinforcement learning refined with physical feedback to ensure robustness and stability. Extensive evaluation in simulation and real-world experiments demonstrates that Humanoid-LLA achieves superior generalization to novel language commands and diverse motion generation while maintaining high physical fidelity.

1 INTRODUCTION

Enabling humanoid robots to follow free-form language instructions while maintaining physically stable whole-body behaviors remains a critical challenge for embodied intelligence. Recent progress in large language models (LLMs) (Shao et al., 2024; Yang et al., 2025a) and vision-language-action models (VLA) (Kim et al., 2024; Bjorck et al., 2025; Xu et al., 2024b; Xue et al., 2025; Ding et al., 2025) has begun to bridge foundation models with embodied systems. Despite VLAs having shown strong results in manipulation tasks, free-form language control for humanoid robots remains challenging due to the high degree of freedom and complex dynamics inherent in humanoid robots, specifically: (i) paired data of language instructions and physically executable whole-body motions for humanoids is expensive and difficult to collect at scale, and prior pipelines further discard a large fraction of human text–motion data due to non-executability on a specific humanoid; (ii) whole-body humanoid motions are highly sensitive to physical instability, where small errors can be amplified through contact dynamics and lead to falls.

Recent pioneering works have begun to explore simple language-conditioned whole-body control. These efforts generally fall into two categories. The first and more established paradigm relies on motion mimicking frameworks (He et al., 2025b; Yue et al., 2025; Xu et al., 2024a). In this pipeline, human motions are first generated from text using off-the-shelf models (Tevet et al., 2023; Zhang et al., 2023; Jiang et al., 2024; Team, 2025) and subsequently retargeted to humanoid robots. However, directly coupling generation with retargeting (Araujo et al., 2025; Yang et al., 2025b) introduces significant challenges. A key issue is that text-to-human generators, trained solely on human motion distributions, often produce outputs that violate humanoid dynamics, necessitating labor-intensive curation for physical feasibility. The second paradigm utilizes distillation frameworks(Shao et al., 2025; Xue et al., 2025), which transfer knowledge from a privileged tracking controller to a text-conditioned policy. While this method achieves strong physical fidelity, it typically compresses both semantic understanding and motor control into a single variational model. This compression often weakens the model's grounding in language and blurs precise action selection. Critically, both paradigms are fundamentally limited to mapping simple instructions to motions. They lack the deep

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.

comprehension, reasoning, and generative capabilities of modern large language models, failing to translate complex, high-level linguistic commands into nuanced and contextually appropriate whole-body behaviors.

To address these limitations, we introduce Humanoid-LLA, a Large Language-Action model designed for free-form text-to-humanoid whole-body control. Our model directly maps expressive natural language instructions into executable humanoid actions, leveraging its strong reasoning and analytical capabilities to interpret and decompose high-level commands.

To overcome the fundamental challenge of data scarcity in humanoid whole-body control, we propose building a unified human–humanoid discrete motion vocabulary that maximizes the utility of existing motion data. Importantly, this "data scarcity" is not about the lack of human text–motion data, but about the limited amount of paired trajectories that are physically executable on a specific humanoid: prior pipelines typically filter out a substantial fraction of human motions as non-executable or rely on costly curation, shrinking motion–language coverage and weakening broad language grounding. Our unified vocabulary enables leveraging the full human corpus to learn language–motion alignment, while using the humanoid-executable subset to learn physical grounding. Specifically, we construct this shared vocabulary through joint quantization of paired human motions and their retargeted humanoid counterparts. This process enforces bidirectional cross-embodiment reconstruction, ensuring that the same discrete token consistently represents the same motion primitive across both human and humanoid embodiments.

Moreover, to enhance physical stability for robust execution, we introduce two key components: (1) A vocabulary-directed controller, distilled from a privileged tracking teacher policy, serves as the critical bridge that converts an LLM-generated vocabulary sequence into stable motor execution. This controller provides implicit vocabulary tracking, eliminating the need for online retargeting at inference time. (2) Reinforcement learning fine-tuning guided by humanoid physical feedback. This process involves executing proposed vocabulary sequences in simulation and using a physical reward signal to steer the model's output toward the humanoid robot's feasible motion manifold. Extensive evaluations in both simulation and real-world environments demonstrate compelling language generalization capabilities while maintaining high physical fidelity.

We summarize our core contributions as follows:

- We present Humanoid-LLA, a Large Language-Action model for free-form text-to-humanoid whole-body control, mapping expressive natural-language commands to physically executable humanoid actions.
- We propose a unified human–humanoid motion vocabulary learned via cross-embodiment tokenization, providing a shared discrete interface for language modeling, humanoid control, and cross-embodiment transfer, thereby mitigating embodiment mismatch and data scarcity in whole-body humanoid control.
- To bridge the gap between high-level instruction understanding and physically constrained execution, we propose a systematic two-stage fine-tuning framework. By sequentially combining supervised motion CoT learning and RL with physical feedback, Humanoid-LLA acquires a unified capability for semantic reasoning and physical stability, enabling robust generalization under real-world dynamics.

2 Related Work

Text-to-Motion Generation. Text-to-motion work typically treats motion synthesis as conditional sequence modeling, producing long-horizon pose trajectories from language. Diffusion-based methods (Tevet et al., 2023; Chen et al., 2023; Zhang et al., 2024; Karunratanakul et al., 2023) yield diverse and high-quality motions but remain expensive and difficult to steer, while GPT-based generators (Zhang et al., 2023; Jiang et al., 2024; Ouyang et al., 2025) improve efficiency and temporal consistency yet are prone to discretization and representation artifacts. Recent efforts (Yuan et al., 2023; Serifi et al., 2024; Han et al., 2025) inject physical priors by projecting kinematic outputs into feasible states or by training with reward surrogates and RL controllers, but the resulting motions are still often generated in a human-centric space and may not be directly executable on a humanoid. In contrast, we aim to learn language–motion grounding in a unified human–humanoid discrete

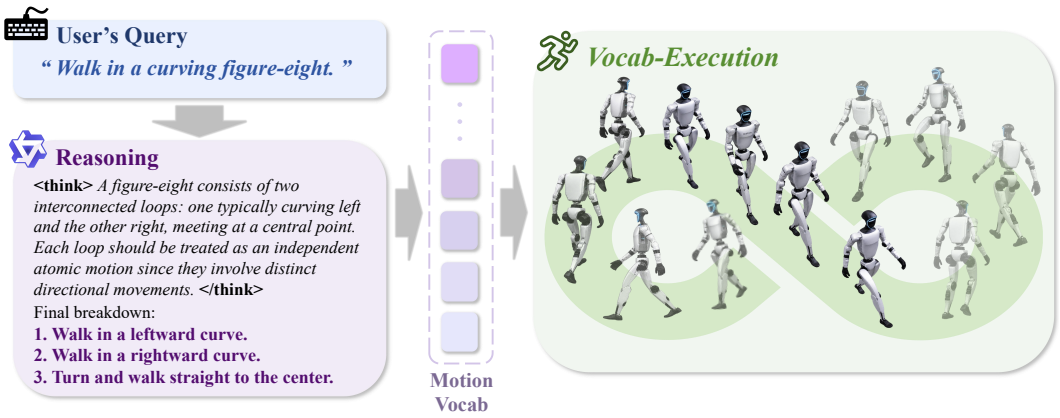


Figure 1: **An illustration of Humanoid-LLA.** Given a high-level and abstract command, Humanoid-LLA first uses natural language to decompose the task into several sub-tasks, and then a downstream controller is responsible for sequential execution, yielding semantically aligned and physically faithful whole-body behaviors.

vocabulary, so that scaling with abundant human data does not decouple generation from downstream humanoid feasibility.

Physics-Based Control with Language Interface. To correct kinematic artifacts such as foot sliding and implausible contacts, physics-based animation trains controllers in simulation to enforce dynamical consistency, enabling robust imitation and skill composition (Peng et al., 2018; 2021; 2022; Tessler et al., 2024). However, specifying rich behaviors purely via low-level rewards or task-specific controllers can be cumbersome, motivating language-conditioned control of physically simulated agents (Juravsky et al., 2022; 2024). Recent systems combine linguistic interfaces with motion codebooks, physics-aware imitation, or closed-loop plan-and-imitate schemes (Yao et al., 2024; Truong et al., 2024; Tevet et al., 2024; Wu et al., 2025), pointing toward closed-loop frameworks that couple semantic intent with physical execution. This work is inspired by this direction but differs by introducing a unified human–humanoid vocabulary interface and a vocab-directed controller, enabling token-level execution as a stable bridge from language generation to real-world physical execution.

Language-Conditioned Humanoid Whole-Body Control. Progress in humanoid control has improved retargeting, teleoperation, and sim-to-real robustness (Luo et al., 2023; Araujo et al., 2025; Yang et al., 2025b; He et al., 2025b; Ze et al., 2025a;b; Ben et al., 2025; Li et al., 2025a;b; He et al., 2024; Cheng et al., 2024; Ji et al., 2024; He et al., 2025c;a; Liao et al., 2025; Yin et al., 2025a; Luo et al., 2025), yet language-conditioned whole-body control (Shi et al., 2025; Shao et al., 2025; Yue et al., 2025; Li et al., 2025c; Wang et al., 2025; Jiang et al., 2025a; Li et al., 2026; Luo et al., 2025; Xie et al., 2026; Bao et al., 2026; Yuan et al., 2026; Cao et al., 2026) still faces a semantic-to-physical gap. Early explorations such as Harmon (Jiang et al., 2025b), UH-1 (Mao et al., 2025), and ALMI (Shi et al., 2025) improve text–humanoid motion corpora but struggle with real-world deployment. LangWBC (Shao et al., 2025) boosts physical fidelity at the cost of language generalization, and RLPF (Yue et al., 2025) fine-tunes an LLM-based generator with physical feedback, but operates in human-motion space and can restrict diversity due to conservative binary reward.

More recently, RoboGhost (Li et al., 2025c) adopts retargeting-free latent-driven control, and SENTINEL (Wang et al., 2025) trains an end-to-end language–action model on a large tracking-built dataset. UniAct (Jiang et al., 2025a) uses shared discrete quantization for multi-modal motion generation. However, these methods largely decouple generation from execution, offering limited dynamic feedback to the generator, and underexplore complex commands requiring semantic decomposition. In contrast, we use simulation-based rewards to realize closed-loop physics-aware token generation. Additionally, our unified motion vocabulary enables cross-embodiment transfer from human motion data, with extensive real-world experiments on Unitree G1 and Booster T1 demonstrating the paradigm's effectiveness and generalizability.

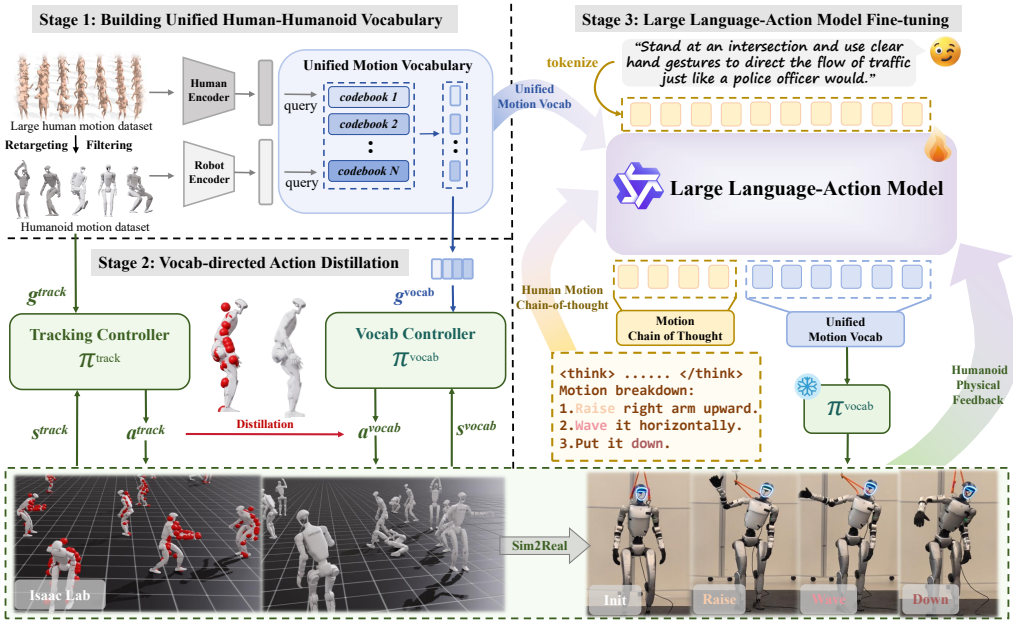


Figure 2: **Overview of Humanoid-LLA.** In stage one, we build a unified motion vocabulary leveraging a large-scale paired human and humanoid motion dataset. With a kinematic humanoid motion goal and its corresponding vocab retrieval, we distill a vocab-directed humanoid student controller from a teacher tracking controller. The first two stages enable stage three to acquire various humanoid feedback directly from physical simulation without decoding, making our LLA enhanced with high physical fidelity and language generalization.

3 METHOD

Our framework, as shown in Fig. 2, consists of three tightly connected components: building unified human–humanoid motion vocabulary (Sec. 3.1), distilling vocabulary-directed policy (Sec. 3.2), and fine-tuning large language-action model (Sec. 3.3). The first two components serve as essential prerequisites that make the integrated reasoning in the third component possible. Next, we introduce each component, highlighting its role within the overall framework.

3.1 UNIFIED HUMAN-HUMANOID VOCABULARY

Humanoid Motion Canonicalization. To enable cross-embodiment alignment, we canonicalize humanoid motions into a structured, root-centered representation that is compatible with the commonly used human motion parameterization (Guo et al., 2022). The representation aggregates joint configuration together with essential kinematic signals, and is normalized to a consistent coordinate frame for learning a shared vocabulary.

Implicit Partitioning Tokenization. We aim to learn a unified tokenizer that maps human and retargeted humanoid motions into the same discrete vocabulary, ensuring that identical tokens carry consistent semantics across embodiments. This vocabulary serves as a compact interface for language-model reasoning and enables scalable training with abundant human motion data.

A discrete interface, however, may overcompress fine-grained kinematics required by stable whole-body control. To mitigate this information bottleneck and avoid manual body-part-based partitioning of motion representations, we adopt VQ-VAE (Van Den Oord et al., 2017) with implicit partitioning (Ma et al., 2025), where each latent vector is factorized into sub-blocks and quantized by multiple sub-codebooks. This factorization increases representational bandwidth without requiring a single giant codebook, thereby preserving motion details that are critical for stable execution.

Cross-embodiment Optimization. We optimize the dual-branch VQ-VAE by combining intra-modal and cross-modal reconstruction objectives. A sequence of human motion $\mathbf{m} \in \mathbb{R}^{T\times d_h}$ and humanoid motion $\mathbf{m} \in \mathbb{R}^{T\times d_r}$ are first encoded into latent features $\mathbf{z}^{\text{h}} = \mathcal{E}_{\text{human}}(\mathbf{m}^{\text{h}})$ and $\mathbf{z}^{\text{r}} = \mathcal{E}_{\text{robot}}(\mathbf{m}^{\text{r}})$, which are partitioned into sub-blocks and quantized by multiple codebooks to yield discrete tokens $\hat{\mathbf{z}}^{\text{h}}$ and $\hat{\mathbf{z}}^{\text{r}}$. These tokens are then decoded back to the motion space by modality-specific decoders $\mathcal{D}_{\text{human}}$ and $\mathcal{D}_{\text{robot}}$, producing both self-reconstructions ($\hat{\mathbf{m}}^{\text{h}}$, $\hat{\mathbf{m}}^{\text{r}}$) and cross-reconstructions ($\hat{\mathbf{m}}^{\text{r}\leftarrow\text{h}}$, $\hat{\mathbf{m}}^{\text{h}\leftarrow\text{r}}$). Beyond standard self-reconstruction within each modality (Zhao et al., 2025), we additionally enforce cross-modal reconstruction, such that a token obtained from either modality is decoded into the same motion primitive, which is crucial for achieving semantically unified tokenization. The training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{intra}} + \alpha\mathcal{L}_{\text{commit}} + \beta\mathcal{L}_{\text{cross}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \qquad \mathcal{L}_{\text{reg}} = \|z^h - z^r\|_1. \tag{1}$$

where $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{cross}}$ are the within-embodiment and cross-embodiment reconstruction losses, respectively; $\mathcal{L}_{\text{commit}}$ is the VQ commitment loss. For temporally aligned human–humanoid pairs, $\mathcal{L}_{\text{reg}}$ brings their pre-quantization latents closer.

3.2 Vocabulary-directed Humanoid Action Distillation

With unified motion vocabulary in Sec. 3.1, we next bridge the gap between kinematic motion primitives and physical control through a vocabulary-directed distillation process. Following the teacher–student paradigm used in recent whole-body controllers (He et al., 2025b; Yin et al., 2025a; Tessler et al., 2024; Yin et al., 2025b), we train a privileged teacher policy to track continuous humanoid-retargeted motions with high fidelity and then distill its behavior into a vocabulary-directed student policy that relies on motion tokens. This stage shifts the control input from dense reference trajectories to the compact motion language of tokens, enabling the humanoid to execute token sequences output by the language model in Sec. 3.3.

Fully-constrained Teacher Controller. We follow the goal-conditioned reinforcement learning framework to train a fully-constrained teacher tracking policy π^{track} that tracks dense humanoid-retargeted reference states. At timestep t, the controller observes humanoid proprioception $\mathbf{s}_t$ and a goal state $\mathbf{g}_t^{\text{track}}$ comprising kinematic reference motion, and computes target joint positions $\mathbf{a}_t$ for the PD controller.

The teacher proprioception $\mathbf{s}_t$ consists of the current root linear velocity $\dot{\mathbf{p}}_t^{\text{root}} \in \mathbb{R}^3$, root angular velocity $\omega_t^{\text{root}} \in \mathbb{R}^3$, joint positions $\mathbf{q}_t \in \mathbb{R}^{n_j}$, joint velocities $\dot{\mathbf{q}}_\mathbf{t} \in \mathbb{R}^{n_j}$ and the previous action history $\mathbf{a}_{t-1} \in \mathbb{R}^{n_j}$ with respect to the robot's local coordinate frame:

$$\mathbf{s}_t = \left[\dot{\mathbf{p}}_t^{\text{root}},\ \omega_t^{\text{root}},\ \mathbf{q}_t,\ \dot{\mathbf{q}}_t,\ \mathbf{a}_{t-1}\right]. \tag{2}$$

And for tracking goal observation $\mathbf{g}_t^{\text{track}}$, we track relative body pose instead of absolute poses following a previous tracking framework (Liao et al., 2025):

$$\mathbf{g}_t^{\text{track}} = \left[\hat{\mathbf{q}}_{t+1},\ \hat{\dot{\mathbf{q}}}_{t+1},\ \hat{\mathbf{p}}_{t+1}^{\text{root}} - \mathbf{p}_t^{\text{root}},\ \hat{\theta}_{t+1}^{\text{root}} \ominus \theta_t^{\text{root}}\right], \tag{3}$$

where $\ominus$ denotes the difference between two rotations. The policy action $\mathbf{a}_t$ is the normalized robot target joint positions, which are residual targets for nominal joint configuration.

For policy training, Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm is used to maximize the accumulated reward $J(\pi^{\text{track}}) = \mathbb{E}_{\pi^{\text{track}}}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\right]$. We design the reward r_t as a weighted sum of task rewards, regularization, and penalty. Details can be found in the Appendix.

Vocabulary-directed Student Controller. After the fully-constrained teacher controller is trained, we distill π^{track} into a vocabulary-directed student policy. Let the unified tokenizer (Sec. 3.1) provide a motion vocab window $\hat{\mathbf{z}}_{1:T}^{\text{vocab}}$, we aim to train a student policy π^{vocab} that can generate full body actions satisfying these given motion vocabulary commands. To solve this ambiguity, we follow prior work (Tessler et al., 2024; 2025) and model π^{vocab} as a Conditional Variational Autoencoder (CVAE) (Kingma & Welling, 2013) consisting of a vocabulary prior ρ, a residual encoder $\mathcal{E}$ and an action decoder $\mathcal{D}$. At timestep t, the motion vocab observation of the student controller is:

$$\mathbf{g}_t^{\text{vocab}} = \left[\mathcal{M}(\mathbf{g}_t^{\text{track}}),\ \hat{\mathbf{z}}_t^{\text{vocab}}\right], \tag{4}$$

where $\mathcal{M}(\cdot)$ is a random masking function and $\hat{\mathbf{z}}_t^{\text{vocab}}$ is the current motion vocabulary in Sec. 3.1. The vocabulary prior is modeled as a Gaussian distribution over latents given the observed vocab constraints:

$$\rho(z_t|\mathbf{s}_t, \mathbf{g}_t^{\text{vocab}}) = \mathcal{N}\left(\mu^\rho(\mathbf{s}_t, \mathbf{g}_t^{\text{vocab}}),\ \sigma^\rho(\mathbf{s}_t, \mathbf{g}_t^{\text{vocab}})\right). \tag{5}$$

The encoder $\mathcal{E}$ is modeled as a residual to the prior that outputs a latent distribution given the full-constraint teacher observation $\mathbf{g}_t^{\text{track}}$ (Yao et al., 2022):

$$\mathcal{E}(z_t \mid \mathbf{s}_t, \mathbf{g}_t^{\text{track}}) = \mathcal{N}\Big(\mu^\rho(\mathbf{s}_t, \mathbf{g}_t^{\text{vocab}}) + \mu^{\mathcal{E}}(\mathbf{s}_t, \mathbf{g}_t^{\text{track}}),\ \sigma^{\mathcal{E}}(\mathbf{s}_t, \mathbf{g}_t^{\text{track}})\Big). \tag{6}$$

Based on the Dataset Aggregation (DAgger) algorithm (Ross et al., 2011), we train π^{vocab} from π^{track} with motion vocab labels within the same motion dataset. The training objective is to minimize the difference between reference action and student action as well as the KL divergence between encoder distribution $p_{\mathcal{E}}$ and prior distribution q_ρ:

$$\mathcal{L}_{\pi^{\text{vocab}}} = \|a_t^{\text{track}} - a_t^{\text{vocab}}\|_2^2 + \lambda_{\text{KL}}\Big(p_{\mathcal{E}}(z_t \mid \mathbf{s}_t,\ \mathbf{g}_t^{\text{track}}) \,\Big\|\, q_\rho(z_t \mid \mathbf{s}_t,\ \mathbf{g}_t^{\text{vocab}})\Big), \tag{7}$$

where a_t^{track} is the reference action from π^{track}, a_t^{vocab} is the student action sampled from $\mathcal{D}(a_t^{\text{vocab}}|\mathbf{s}_t, \mathbf{g}_t^{\text{vocab}})$ and λ_{KL} is the hyperparameter for balancing reconstruction and regularization.

3.3 Large Language-Action Model

In this section we show how, building upon Sec. 3.1 and Sec. 3.2, our framework realizes a direct mapping from highly abstract language descriptions to physically executable robot actions without decoding and retargeting during inference. Sec. 3.2 serves as the key intermediate: a low-level controller distilled to follow latent motion tokens, seamlessly linking latent motion token generation and physics-based action execution. The remainder of this section details the training procedure.

Supervised Fine-tuning with Augmented Human Data. To solve the limited expressivity of textual annotations in existing datasets (Guo et al., 2022; Punnakkal et al., 2021; Hwang et al., 2026), prior approaches (Ouyang et al., 2025) have resorted to LLM (Shao et al., 2024) to decompose abstract motion descriptions. However, such methods often cause motion-language misalignment since a single caption may correspond to multiple plausible motions. Inspired by Cui et al. (2024), who demonstrate the effectiveness of visual signal for motion-text understanding, we employ a vision–language model (Bai et al., 2025) with rendered motion sequences as motion context, enabling it to produce more accurate motion CoT. We use motion CoT only for abstract or compositional instructions, and atomic commands are passed through unchanged.

Given the augmented large-scale human motion–text datasets, we formulate motion token generation as an autoregressive, text-conditioned language modeling task, where a motion sequence is represented as a series of discrete tokens from the unified codebook $\mathcal{Z} = \{\langle cb_{i,j}\rangle\}$, with i indexing the sub-codebook and j the token entry. The input is the textual description $\mathbf{w}$, and the supervision target $\mathbf{y} = (y_1, \ldots, y_L)$ is constructed by concatenating motion CoT with the ground-truth motion tokens from the pretrained tokenizer. The model is trained with the standard next-token prediction loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\mathbf{w},\mathbf{y})\sim\mathcal{D}} \sum_{t=1}^{L} \log P_\phi\big(y_t \mid \mathbf{w}, y_{<t}\big), \tag{8}$$

where ϕ are the model parameters. Importantly, we perform SFT for language–motion semantic alignment using the full human text–motion corpus under the unified vocabulary; physical executability is not required at this stage. We only enforce humanoid feasibility in the later stage through our pretrained vocab-controller and physics-based RL fine-tuning on the G1-executable subset.

RL Fine-tuning with Humanoid Feedback. We fine-tune the motion-token generator with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each prompt x, the rollout policy π_{old} samples K candidate sequences. Their scalar rewards are normalized within the group to obtain sequence-level advantages $\tilde{A}_k$. We optimize the token-level clipped objective

$$\mathcal{L}_{\mathrm{GRPO}}(\phi) = -\mathbb{E}_{x\sim\mathcal{D},\, y^{(1:K)}\sim\pi_{\mathrm{old}}}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{T_k}\sum_{t=1}^{T_k}\ell_{k,t}\right] + \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}},$$
$$\ell_{k,t} = \min\Big(\rho_{k,t}\tilde{A}_k,\ \mathrm{clip}_\epsilon(\rho_{k,t})\tilde{A}_k\Big), \qquad \rho_{k,t} = \frac{\pi_\phi(y_t^{(k)} \mid h_{k,t})}{\pi_{\mathrm{old}}(y_t^{(k)} \mid h_{k,t})}. \tag{9}$$

Here $T_k = |y^{(k)}|$, $h_{k,t} = (x, y^{(k)}_{<t})$, and $\mathrm{clip}_\epsilon(z)$ clips z to $[1-\epsilon, 1+\epsilon]$. The group-relative advantage is $\tilde{A}_k = (R_k - \bar{R})/(\sigma_R + \delta)$, where $\bar{R}$ and σ_R are the mean and standard deviation of the K sequence rewards, and $\delta > 0$ ensures numerical stability. The KL term $\mathcal{L}_{\mathrm{KL}}$ keeps π_ϕ close to the fixed SFT reference policy π_{ref}.

Unlike prior work that emphasizes kinematic fidelity (Ouyang et al., 2025; Yue et al., 2025), we stress the importance of dynamics-level consistency for real-world deployment. RLPF (Yue et al., 2025) employs a binary simulator-tracking reward, which ensures executability but often reduces motion diversity, as the policy tends to favor conservative behaviors that are easy to track. To address this, we design a reward scheme that combines high-level distributional objectives with low-level simulator-based tracking signals, achieving motions that are both physically robust and expressively varied.

Physical Fidelity Reward Design. The overall reward is a weighted sum of a binary format reward and a continuous physical fidelity reward. The format reward acts as a prerequisite: the model must first learn *how to answer* (i.e., producing valid structured outputs) before it can effectively learn *how to answer well* (i.e., generating physically and semantically aligned motions). Concretely, the format reward checks two requirements: (i) the response must follow a structured template beginning with `<think>...</think>` and followed by `<motion>...</motion>`; and (ii) within the motion segment, motion tokens must appear in cyclic sub-codebook order (`cb0`→`cb1`→...→`cb(N-1)` repeatedly). We define it as

$$r_{\text{format}} = \mathbb{I}\{\text{requirements satisfied}\}. \tag{10}$$

The physical fidelity reward is composed of a distributional term and a tracking term. The distributional reward encourages motion distribution generated by the vocab-controller to match the distribution of physically plausible trajectories and to align semantically with the paired motion descriptions. Using contrastive motion encoder $\phi_{\mathrm{m}}(\cdot)$ and text encoder $\phi_{\mathrm{t}}(\cdot)$ (Guo et al., 2022) trained on physically plausible humanoid datasets, we define distributional reward as:

$$r_{\text{dist}} = \exp\big(-\lambda_m \,\|\phi_{\mathrm{m}}(\mathbf{m}_{\text{gen}}) - \phi_{\mathrm{m}}(\mathbf{m}_{\text{ref}})\|_2\big) + \exp\big(-\lambda_t \,\|\phi_{\mathrm{m}}(\mathbf{m}_{\text{gen}}) - \phi_{\mathrm{t}}(\mathbf{w}_{\text{ref}})\|_2\big), \tag{11}$$

where the two terms collectively measure motion fidelity and semantic fidelity. $\mathbf{m}_{\text{gen}}$, $\mathbf{m}_{\text{ref}}$ and $\mathbf{w}_{\text{ref}}$ represent the motion rolled out in simulation by the vocab-controller, ground-truth motion and paired motion description, respectively. λ_m, λ_t are balancing coefficients.

The tracking reward measures how well a generated token sequence can be executed in simulation by the distilled vocab-controller (Sec. 3.2). We evaluate the simulated rollout with a position reward term r_{pos} and an acceleration reward term r_{acc}:

$$r_{\text{track}} = r_{\text{pos}} + r_{\text{acc}}. \tag{12}$$

Finally, the overall reward is calculated as $r = r_{\text{format}} + r_{\text{dist}} + r_{\text{track}}$. For more details about the Large Language-Action Model design, see Appendix.

4 Experiments

4.1 Experiment Setup

Dataset. We conduct experiments on the text-annotated subset of AMASS (Mahmood et al., 2019; Guo et al., 2022), comprising ∼27K motion sequences. We use `mink` (Zakka, 2024) to retarget human motions to the humanoid and a privileged tracking teacher to filter physically infeasible motions, retaining approximately 70% as a G1-executable subset. Qwen2.5-VL (Bai et al., 2025) generates motion CoT annotations from the raw descriptions and rendered sequences. AMASS provides low-noise mocap data and is widely used in motion generation and humanoid control, supporting comparison across methods; Appendix A.2 gives data-curation details.

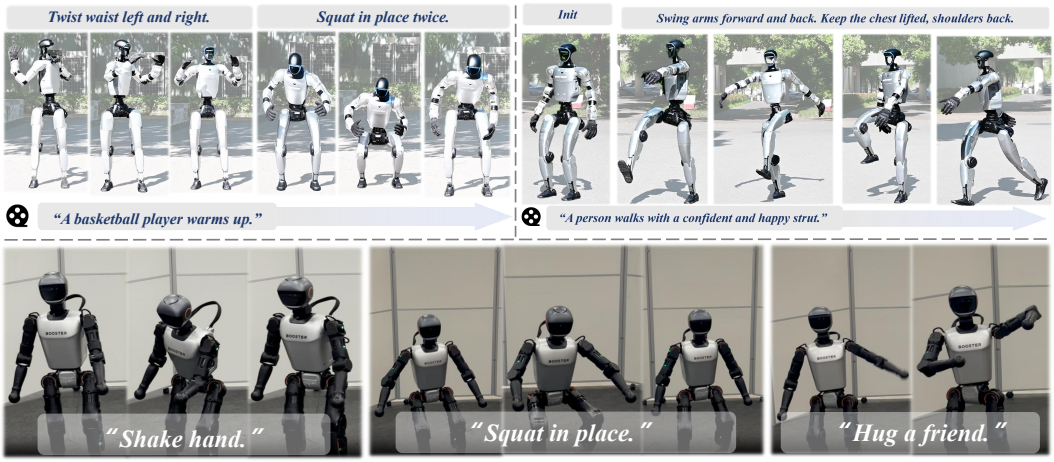


Figure 3: Real-world experiments on Unitree G1 (top) and Booster T1 (bottom). We test both simple instructions and abstract instructions that require decomposition. The results demonstrate that our method simultaneously achieves semantic consistency and physical robustness.

Table 1: Text-to-humanoid results: generation metrics on the left and physics-based metrics on the right. R-Precision is reported at top-3. Arrows indicate the preferred direction.

Methods	FID↓	R-Prec.↑	MM-Dist↓	Div.→
MDM+Retarget (Tevet et al., 2023)	11.759	0.262	6.599	6.419
OmniH2O (He et al., 2025b)	17.159	0.222	8.021	5.868
UH-1 (Mao et al., 2025)	8.682	0.295	5.896	6.749
LangWBC* (Shao et al., 2025)	6.171	0.320	5.587	6.031
UniAct-T2M (Jiang et al., 2025a)	8.513	0.284	6.031	6.514
Humanoid-LLA	**2.626**	**0.447**	**4.911**	**7.122**

Methods	Succ.↑	MPJPE↓	E_{vel} ↓	E_{acc} ↓
OmniH2O (He et al., 2025b)	72.2%	73.43	11.78	10.48
UH-1 (Mao et al., 2025)	68.8%	121.51	16.59	14.80
LangWBC* (Shao et al., 2025)	76.0%	–	–	–
RLPF (Yue et al., 2025)	80.0%	140.00	–	–
UniAct-T2M (Jiang et al., 2025a)	82.5%	60.33	9.15	8.93
Humanoid-LLA	**87.6%**	**56.43**	**8.92**	**8.74**

Baselines. We compare MDM+Retarget (Tevet et al., 2023), OmniH2O (He et al., 2025b), UH-1 (Mao et al., 2025), LangWBC (Shao et al., 2025), RLPF (Yue et al., 2025), and UniAct-T2M (Jiang et al., 2025a). They span kinematic generation and retargeting, tracking-based execution, direct humanoid generation, language-conditioned control, and physics-aware fine-tuning. Appendix A.1.4 gives implementation details.

Evaluation Metrics. Most prior work on text-to-humanoid motion generation (Mao et al., 2025; Shao et al., 2025; Shi et al., 2025; Yue et al., 2025) reports either low-level physics tracking metrics or human-motion generation metrics, leaving no unified protocol directly defined on humanoid robots. To fill this gap, we design an evaluation that combines physics-based tracking measures with distributional generation metrics computed in humanoid motion space. These two perspectives jointly capture executability, distributional fidelity, motion–language alignment, and diversity, thus discouraging models from producing only simple, easily executable motions at the expense of expressiveness. For the generation side, we report FID to measure distributional similarity against a physical humanoid motion set obtained by a goal-conditioned tracking policy (i.e., teacher controller in Sec. 3.2), MM-Dist and R-Precision to assess motion-language alignment, and Diversity (Div.) to evaluate variability. For the physics side, we measure success rate (Succ.), mean per-joint position error MPJPE (mm), velocity error E_{vel} (mm/frame), and acceleration error E_{acc} (mm/frame^2). Refer to Appendix for details.

4.2 Text-to-Humanoid Evaluation

Tab. 1 reveals trade-offs between generation quality and simulated physical execution among baselines. MDM+Retarget (Tevet et al., 2023) transfers expressive, diverse human motions through kinematic retargeting but lacks physical fidelity. OmniH2O (He et al., 2025b) adds an imitation policy for feasible trajectories, yet human–robot action-space mismatch still causes frequent tracking failures. UH-1 (Mao et al., 2025) decodes from a robot-space latent manifold trained on robot trajectories but remains limited for real-world deployment. LangWBC (Shao et al., 2025) conditions jointly on

Table 2: Ablation results of CoT reasoning, RL fine-tuning, and physical reward terms.

Methods	FID↓	R-Precision↑	MM-Dist↓	Div.→	Succ.↑	MPJPE↓	E_{vel} ↓	E_{acc} ↓
Humanoid-LLA w/o CoT	10.423	0.270	6.222	6.405	64.90%	90.43	14.11	11.23
Humanoid-LLA w/o RLFT	5.132	0.331	5.443	6.668	68.64%	78.31	12.12	10.01
Humanoid-LLA w/o r_{dist}	4.597	0.342	5.401	6.892	85.33%	61.27	9.31	9.02
Humanoid-LLA w/o r_{track}	**2.578**	0.439	5.013	7.007	76.72%	66.42	10.89	9.77
Humanoid-LLA (Ours)	2.626	**0.447**	**4.911**	**7.122**	**87.6%**	**56.43**	**8.92**	**8.74**

language and control, yielding strong low-level execution but weaker motion–language alignment. RLPF (Yue et al., 2025) uses physical feedback to constrain feasibility, although optimizing motion distributions in human space limits humanoid alignment. UniAct-T2M (Jiang et al., 2025a) generates motion tokens autoregressively without motion CoT supervision or humanoid-feedback RL fine-tuning. Humanoid-LLA instead couples LLM-generated tokens with a vocabulary-directed controller and humanoid-feedback fine-tuning, preserving diversity and expressiveness while substantially improving physical fidelity. This yields consistent gains on both generation and tracking metrics over the evaluated baselines. Implementation details are in Appendix A.1.4.

4.3 Free-form Instructions Evaluation

We additionally evaluate two free-form sets in simulation (Tab. 3): 100 LLM-augmented prompts and 100 human-written prompts from SnapMoGen (Hwang et al., 2026). Humanoid-LLA leads both sets in success and MM-Dist; Appendix A.2.3 gives the construction and analysis.

Table 3: Free-form instruction evaluation in simulation.

LLM-augmented prompts			**Human-written prompts**		
Method	Succ.↑	MM-Dist↓	Method	Succ.↑	MM-Dist↓
LangWBC* (Shao et al., 2025)	65%	10.34	LangWBC* (Shao et al., 2025)	61%	10.89
FRoM-W1 (Li et al., 2026)	66%	8.58	FRoM-W1 (Li et al., 2026)	60%	9.32
TextOp (Xie et al., 2026)	63%	8.63	TextOp (Xie et al., 2026)	66%	9.04
Humanoid-LLA w/o CoT	68%	8.77	Humanoid-LLA w/o CoT	65%	9.21
Humanoid-LLA (Ours)	**74%**	**6.12**	Humanoid-LLA (Ours)	**78%**	**6.44**

4.4 Ablation Studies

We ablate motion CoT, RL fine-tuning (RLFT), and both physical rewards. Without CoT, the model uses raw motion descriptions; without RLFT, it reverts to SFT-only. Removing r_{dist} or r_{track} retains the other reward. Tab. 2 shows that removing CoT worsens FID (2.626 → 10.423), R-Precision (0.447 → 0.270), and success (87.6% → 64.90%), indicating that intermediate reasoning helps map complex instructions to motion tokens. RLFT improves success (68.64% → 87.6%), MPJPE (78.31 → 56.43), and FID (5.132 → 2.626) over SFT-only, indicating that physical feedback reduces the mismatch between generated motions and feasible humanoid behavior. Reward ablations separate their roles: removing r_{dist} mainly worsens FID and R-Precision, whereas removing r_{track} lowers success to 76.72% and increases tracking error; combining both yields the best trade-off.

5 Conclusion

In this work, we present Humanoid-LLA, a unified framework for free-form language humanoid control that bridges high-level language commands and whole-body execution. Specifically, Humanoid-LLA introduces a unified motion vocabulary that aligns human and humanoid motion primitives, effectively bridging large language models and a whole-body controller. By augmenting large-scale human-motion datasets with vision-language model-generated annotations and fine-tuning with humanoid feedback in simulation, our model achieves enhanced language generalization and physical feasibility at execution. On the reported benchmark, the method improves generation and simulated tracking metrics relative to the evaluated baselines. Hardware demonstrations on G1 and an adapted T1 controller establish feasibility but also expose failures on dynamic skills. Longer-horizon planning, richer multimodal context, and more dynamics-aware retargeting remain important next steps.

AI USE STATEMENT

Generative AI tools were used to generate and verify motion-language annotations for dataset augmentation. Specifically, Qwen2.5-VL was used to produce motion CoT annotations from text and rendered motion sequences, and GPT-4o was used to verify annotation consistency and identify hallucinations or formatting errors. Generative AI tools were also used to edit the manuscript for clarity and concision. All AI-assisted outputs used in the work were reviewed by the authors, who take responsibility for the final content.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. This work uses publicly available research datasets and in-lab robot experiments conducted under standard safety protocols, including emergency-stop procedures, safety clearance zones, and supervised operation. We followed the applicable dataset licenses and terms of use and did not collect personally identifiable information or sensitive biometric data. We additionally conducted a small-scale perceptual user study with 25 participants within our institution to evaluate the generated robot motions. The study involved only subjective comparisons of motion videos and did not collect personally identifiable or sensitive information. Participation was voluntary. The study was conducted in accordance with applicable institutional policies and standard research practices.

REPRODUCIBILITY STATEMENT

The main paper describes the proposed method and reports the principal experimental results. The appendix provides further information on implementation, training, data preparation, evaluation, and experimental settings to support reproduction. We plan to release relevant code and supporting materials when possible, subject to the licenses and access restrictions that apply to third-party data and hardware.

REFERENCES

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jiacheng Bao, Haoran Yang, Yucheng Xin, Junhong Liu, Yuecheng Xu, Han Liang, Pengfei Han, Xiaoguang Ma, Dong Wang, and Bin Zhao. Phygile: Physics-prefix guided motion generation for agile general humanoid motion tracking. *arXiv preprint arXiv:2603.19305*, 2026.

Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Jianuo Cao, Yuxin Chen, Yuzhen Song, Masayoshi Tomizuka, Chenran Li, and Thomas Tian. Texedo : Test time scaling for controller-aware language-conditioned humanoid motion generation, 2026. URL https://arxiv.org/abs/2606.22998.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18000–18010, 2023.

Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.

Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

Thomas Flayols, Andrea Del Prete, Patrick Wensing, Alexis Mifsud, Mehdi Benallegue, and Olivier Stasse. Experimental evaluation of simple estimators for humanoid robots. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 889–895. IEEE, 2017.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161, 2022.

Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2218–2227. IEEE, 2025.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8944–8951. IEEE, 2024.

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025a.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning*, pp. 1516–1540. PMLR, 2025b.

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9989–9996. IEEE, 2025c.

Inwoo Hwang, Jian Wang, Bing Zhou, et al. Snapmogen: Human motion generation from expressive texts. *Advances in Neural Information Processing Systems*, 38:99939–99955, 2026.

Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

Nan Jiang, Zimo He, Wanhe Yu, Lexi Pang, Yunhao Li, Hongjie Li, Jieming Cui, Yuhan Li, Yizhou Wang, Yixin Zhu, et al. Uniact: Unified motion generation and action streaming for humanoid robots. *arXiv preprint arXiv:2512.24321*, 2025a.

Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. In *Conference on Robot Learning*, pp. 3015–3026. PMLR, 2025b.

Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Superpadl: Scaling language-directed physics-based control with progressive supervised distillation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jialong Li, Xuxin Cheng, Tianshu Huang, Shiqi Yang, Rizhao Qiu, and Xiaolong Wang. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. *Robotics: Science and Systems 2025*, 2025a.

Peng Li, Zihan Zhuang, Yangfan Gao, Yi Dong, Sixian Li, Changhao Jiang, Shihan Dou, Zhiheng Xi, Enyu Zhou, Jixuan Huang, et al. From-w1: Towards general humanoid whole-body control with language instructions. *arXiv preprint arXiv:2601.12799*, 2026.

Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks, 2025b.

Zhe Li, Cheng Chi, Yangyang Wei, Boan Zhu, Yibo Peng, Tao Huang, Pengwei Wang, Zhongyuan Wang, Shanghang Zhang, and Chang Xu. From language to locomotion: Retargeting-free humanoid control via motion latent guidance. *arXiv preprint arXiv:2510.14952*, 2025c.

Qiayuan Liao, Zhongyu Li, Akshay Thirugnanam, Jun Zeng, and Koushil Sreenath. Walking in narrow spaces: Safety-critical locomotion control for quadrupedal robots with duality-based optimization. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2723–2730. IEEE, 2023.

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv e-prints*, pp. arXiv–2508, 2025.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

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, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pp. 5442–5451, October 2019.

Jiageng Mao, Siheng Zhao, Siqi Song, Chuye Hong, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, and Yue Wang. Universal humanoid robot pose learning from internet human videos. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pp. 1–8, 2025. doi: 10.1109/Humanoids65713.2025.11203143.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.

Runqi Ouyang, Haoyun Li, Zhenyuan Zhang, Xiaofeng Wang, Zheng Zhu, Guan Huang, and Xingang Wang. Motion-r1: Chain-of-thought reasoning and reinforcement learning for human motion generation. *arXiv preprint arXiv:2506.10353*, 2025.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 722–731, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a. URL https://arxiv.org/abs/2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021b.

Unitree Robotics. Unitree g1 humanoid robot. https://www.unitree.com/g1, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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

Clemens Schwarke, Mayank Mittal, Nikita Rudin, David Hoeller, and Marco Hutter. Rsl-rl: A learning library for robotics research. *arXiv preprint arXiv:2509.10771*, 2025.

Agon Serifi, Ruben Grandia, Espen Knoop, Markus Gross, and Moritz Bächer. Robot motion diffusion model: Motion generation for robotic characters. In *SIGGRAPH asia 2024 conference papers*, pp. 1–9, 2024.

Yiyang Shao, Xiaoyu Huang, Bike Zhang, Qiayuan Liao, Yuman Gao, Yufeng Chi, Zhongyu Li, Sophia Shao, and Koushil Sreenath. Langwbc: Language-directed humanoid whole-body control via end-to-end learning. *arXiv preprint arXiv:2504.21738*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiyuan Shi, Xinzhe Liu, Dewei Wang, Ouyang Lu, Sören Schwertfeger, Fuchun Sun, Chenjia Bai, and Xuelong Li. Adversarial locomotion and motion imitation for humanoid policy learning. In *Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2504.14305.

Tencent Hunyuan 3D Digital Human Team. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024.

Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. Maskedmanipulator: Versatile whole-body control for loco-manipulation. *arXiv preprint arXiv:2505.19086*, 2025.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *The Thirteenth International Conference on Learning Representations*, 2024.

Takara Everest Truong, Michael Piseno, Zhaoming Xie, and Karen Liu. Pdp: Physics-based character animation via diffusion policy. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–10, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Yuxuan Wang, Haobin Jiang, Shiqing Yao, Ziluo Ding, and Zongqing Lu. Sentinel: A fully end-to-end language-action model for humanoid whole body control. *arXiv preprint arXiv:2511.19236*, 2025.

Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. Uniphys: Unified planner and controller with diffusion for flexible physics-based character control. *arXiv preprint arXiv:2504.12540*, 2025.

Weiji Xie, Jiakun Zheng, Jinrui Han, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Textop: Real-time interactive text-driven humanoid robot motion generation and control. *arXiv preprint arXiv:2602.07439*, 2026.

Shusheng Xu, Huaijie Wang, Yutao Ouyang, Jiaxuan Gao, Zhiyu Mei, Chao Yu, and Yi Wu. Lagoon: Language-guided motion control. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9743–9750. IEEE, 2024a.

Xinyu Xu, Yizheng Zhang, Yong-Lu Li, Lei Han, and Cewu Lu. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *Advances in Neural Information Processing Systems*, 37:18633–18659, 2024b.

Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, et al. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025b.

Heyuan Yao, Zhenhua Song, Baoquan Chen, and Libin Liu. Controlvae: Model-based learning of generative controllers for physics-based characters. *ACM Transactions on Graphics (TOG)*, 41(6): 1–16, 2022.

Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Transactions on Graphics (TOG)*, 43(4):1–21, 2024.

Kangning Yin, Weishuai Zeng, Ke Fan, Zirui Wang, Qiang Zhang, Zheng Tian, Jingbo Wang, Jiangmiao Pang, and Weinan Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *arXiv preprint arXiv:2507.07356*, 2025a.

Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C. Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025b.

Xichen Yuan, Zhe Li, Bofan Lyu, Kuangji Zuo, Yanshuo Lu, Gen Li, and Jianfei Yang. Roboforge: Physically optimized text-guided whole-body locomotion for humanoids. *arXiv preprint arXiv:2603.17927*, 2026.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16010–16021, 2023.

Junpeng Yue, Zepeng Wang, Yuxuan Wang, Weishuai Zeng, Jiangxing Wang, Xinrun Xu, Yu Zhang, Sipeng Zheng, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion models with humanoid control. *arXiv preprint arXiv:2506.12769*, 2025.

Kevin Zakka. Mink: Python inverse kinematics based on mujoco. `https://github.com/kevinzakka/mink`, 2024.

Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025a.

Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025b.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

Haoyu Zhao, Sixu Lin, Qingwei Ben, Minyue Dai, Hao Fei, Jingbo Wang, Hua Zou, and Junting Dong. Smap: Self-supervised motion adaptation for physically plausible humanoid whole-body control. *arXiv preprint arXiv:2505.19463*, 2025.

A Appendix

In this document, we provide expanded technical details, additional experiments, and extended discussions that complement and elaborate on the main paper. Specifically, Sec. A.1 offers a detailed account of our implementation, covering the unified human–humanoid tokenization and codebook construction, the vocabulary-directed control policy with action distillation to low-level joint control, and the Large Language–Action Model interface and training procedure; it also specifies the configurations of baseline methods, evaluation metrics, and the humanoid robot system used in our study. Sec. A.2 presents curated datasets and further analyses of the unified motion vocabulary, a comprehensive suite of experiments and ablations, user studies, and qualitative text-to-humanoid visualizations, and it summarizes the accompanying video. Finally, Sec. A.3 extends the discussions on design choices and limitations, illustrates representative failure cases, and outlines directions for future work. Through this supplementary material, we aim to provide a more complete view of *Humanoid-LLA*, clarify practical nuances for replication, and furnish additional evidence of robustness and versatility.

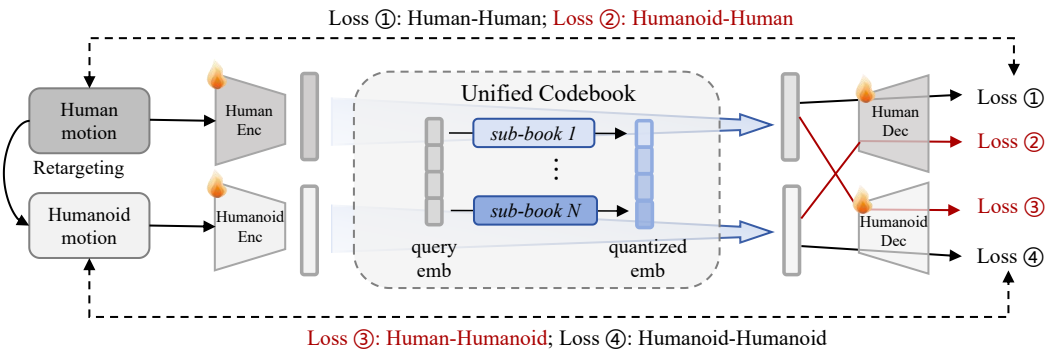


Figure 4: Diagram of detailed unified tokenizer architecture and training procedure.

A.1 Implementation Details

In this section, we provide additional implementation details that complement the main paper, covering (i) the unified human–humanoid motion vocabulary and tokenizer, (ii) vocabulary-directed action distillation and control, (iii) large language–action model training, (iv) baseline implementations and evaluation metrics, and (v) the robot system setup on Unitree G1. Unless otherwise stated, all models are implemented in PyTorch and trained with the GPU configurations specified in the corresponding subsections. To foster reproducibility and to facilitate follow-up research on language-conditioned humanoid control, we commit to releasing our codebase, configuration files, and pretrained checkpoints, as well as scripts for motion retargeting, metric evaluation, and sim-to-real deployment, upon publication of this work, subject to dataset and hardware licensing constraints.

A.1.1 Unified Human–Humanoid Motion Vocabulary

Humanoid Motion Canonicalization. We preprocess raw G1 humanoid trajectories with T frames, each frame represented as

$$x_t = \left[p_t \in \mathbb{R}^3,\ q_t \in \mathbb{R}^4,\ d_t \in \mathbb{R}^{29}\right], \qquad t = 0, \ldots, T-1. \tag{13}$$

We thus produce $(T-1)$ frames of 227-dimensional canonicalized humanoid motion representation per sequence (see Algorithm 1).

Tokenizer Training. To enforce embodiment-agnostic tokenization, we adopt a dual-branch VQ-VAE adapted from Zhang et al. (2023) where human and humanoid motions are encoded separately but quantized through shared codebooks. As shown in Fig. 4, given sequences $\mathbf{m}^{\mathrm{h}} \in \mathbb{R}^{T \times d_h}$ and $\mathbf{m}^{\mathrm{r}} \in \mathbb{R}^{T \times d_r}$, modality-specific encoders produce latents $\mathbf{z}^{\mathrm{h}}, \mathbf{z}^{\mathrm{r}}$, which are quantized into tokens $\hat{\mathbf{z}}^{\mathrm{h}}, \hat{\mathbf{z}}^{\mathrm{r}}$. Decoders then reconstruct both intra- and cross-modal motions. The corresponding objectives are

$$\begin{aligned} \mathcal{L}_{\text{intra}} &= \|\mathbf{m}^{\mathrm{h}} - \hat{\mathbf{m}}^{\mathrm{h}}\|_1 + \|\mathbf{m}^{\mathrm{r}} - \hat{\mathbf{m}}^{\mathrm{r}}\|_1, \\ \mathcal{L}_{\text{cross}} &= \|\mathbf{m}^{\mathrm{h}} - \hat{\mathbf{m}}^{\mathrm{h}\leftarrow\mathrm{r}}\|_1 + \|\mathbf{m}^{\mathrm{r}} - \hat{\mathbf{m}}^{\mathrm{r}\leftarrow\mathrm{h}}\|_1. \end{aligned} \tag{14}$$

where cross-reconstruction ensures that shared tokens decode into semantically consistent motions across embodiments. Notably, we also add a small regularization term $\mathcal{L}_{\text{reg}} = \left\|\mathbf{z}^{\mathrm{h}} - \mathbf{z}^{\mathrm{r}}\right\|_1$ to constrain the unified latent space.

Each encoder–decoder is a temporal convolutional network with depth 3, dilation growth rate 3, and downsampling factor 2^2. The latent space has 512 dimensions, evenly partitioned into 8 sub-chunks of 64 dimensions, each quantized by a codebook of size 64. Training is conducted on text-annotated AMASS with batch size 256. We use AdamW (lr=2×10^{-4}, betas $(0.9, 0.99)$, weight decay 10^{-4}) to optimize the tokenizer training. Both training and evaluation are run on a single NVIDIA RTX 4090 GPU.

Algorithm 1 Humanoid Motion Canonicalization

Input: Raw trajectory $\{x_t\}_{t=0}^{T-1}$ with positions p_t, orientations q_t, DoFs d_t
Output: Canonicalized motion representation $\{f_t\}_{t=0}^{T-2}$
Downsample $\{x_t\}$ from 50 Hz to 20 Hz ▷ Resampling to align with human representation
for $t \leftarrow 0$ **to** $T-1$ **do**
    $r_t \leftarrow \mathrm{rotvec}(q_t)$
    $J_t \leftarrow \{j_t^k \in \mathbb{R}^3\}_{k=1}^{32}$ via FK
▷ Calculate forward kinematics for G1
$h_{\text{floor}} \leftarrow \min_{t,k} j_{t,z}^k$
for $t \leftarrow 0$ **to** $T-1$ **do**
    $p_t \leftarrow p_t - [0, 0, h_{\text{floor}}]^\top$
▷ Floor alignment
for $t \leftarrow 0$ **to** $T-1$ **do**
    $p_t \leftarrow p_t - [p_{0,x}, p_{0,y}, 0]^\top$
$a \leftarrow (j_0^{\text{RHip}} - j_0^{\text{LHip}}) + (j_0^{\text{RShoulder}} - j_0^{\text{LShoulder}})$
$\hat{a} \leftarrow a/\|a\|$
$\hat{f}_0 \leftarrow \dfrac{(0,0,1) \times \hat{a}}{\|(0,0,1) \times \hat{a}\|}$
$q_{\text{align}} \leftarrow \mathrm{Quat}(\hat{f}_0 \mapsto +X)$
for $t \leftarrow 0$ **to** $T-1$ **do**
    $p_t \leftarrow \mathrm{Rot}(q_{\text{align}})\, p_t$
    $q_t \leftarrow q_{\text{align}} \otimes q_t$
▷ Root normalization
for $t \leftarrow 0$ **to** $T-2$ **do**
    $\omega_t \leftarrow r_{t+1} - r_t$ ▷ root angular velocity
    $v_t \leftarrow p_{t+1} - p_t$ ▷ root linear velocity
    $\Delta J_t \leftarrow J_{t+1} - J_t$ ▷ joint velocities
    $f_t \leftarrow [\omega_t,\, v_t,\, \mathrm{vec}(J_t),\, d_t,\, \mathrm{vec}(\Delta J_t)]$
▷ Canonicalized motion representation

Table 4: Reward terms for the fully constrained teacher controller.

Term	Weight	Term	Weight
Task rewards $(\exp(-\alpha\|\cdot\|_2))$			
Root position	0.5	Root rotation	0.5
Body position	1.0	Body rotation	1.0
Body lin. vel.	1.0	Body ang. vel.	0.5
DoF position	2.0	DoF velocity	0.2
Penalties			
Torque limits	-1.0	DoF pos. limits	-5.0
DoF vel. limits	-5.0	Self-contact	-0.1
Regularizers			
Lower-body action rate	-0.4	Upper-body action rate	-0.1
Torque	-10^{-4}	DoF acceleration	-10^{-5}

A.1.2 Vocabulary-Directed Action Distillation

Reward Design. As shown in Tab. 4, we train the fully constrained teacher π^{track} with PPO using a composite reward that combines normalized tracking terms (as exponential scores) with lightweight regularization and hard-limit penalties. Tracking targets are defined in the robot's local frame and computed as relative position and orientation error to reduce drift. Unless noted, all errors are normalized so that weights are comparable across terms.

Domain Randomization. To improve robustness under deployment, we adopt a broad range of randomization during policy training, including varying ground friction, joint damping, sensor latency, and external perturbations. With online adaptation of the teacher to randomized conditions, the distilled tokens encode transferable primitives rather than brittle overfits.

Tracking Controller Training. We train the tracking controller π^{track} with on-policy PPO (Schwarke et al., 2025) using online data collection at a $50\,\text{Hz}$ control rate (physics $\mathrm{d}t = 0.005\,\text{s}$, action hold/decimation $= 4$) and $10\,\text{s}$ episodes in NVIDIA Isaac Lab (Mittal et al., 2023). Each iteration collects 24 policy steps per environment ($\approx 0.48\,\text{s}$ of experience) across up to $16{,}384$ parallel environments on a flat plane, yielding up to $24 \times N_{\text{env}}$ transitions per update. To improve robustness, we apply domain randomization at startup (friction/restitution buckets, joint default pose perturbations, and anchor–body CoM shifts) and inject intermittent external pushes during rollouts by directly setting linear/angular velocities at random intervals between 1–3 s (linear $\pm 0.5\,\text{m/s}$ in x/y, $\pm 0.2\,\text{m/s}$ in z; angular $\pm 0.52\,\text{rad/s}$ roll/pitch, $\pm 0.78\,\text{rad/s}$ yaw). Motion-conditioned commands are loaded from trajectories and sampled with an adaptive time-binning curriculum: the motion timeline is discretized into bins whose sampling probabilities are proportional to recent failure rates, smoothed with a short non-causal kernel (kernel size 3, $\lambda = 0.8$) and updated by an exponential moving average ($\alpha = 0.001$) with a small uniform mixture (ratio 0.1); when an episode terminates or a clip ends, time indices are resampled according to this distribution, and root/joint states are jittered and clipped to soft limits before continuing. PPO optimization uses clipping $\epsilon = 0.2$, learning rate 10^{-3} with an adaptive schedule driven by a desired KL of 0.01, 5 learning epochs over 4 mini-batches per update, value loss coefficient 1.0 with value clipping enabled, entropy coefficient 0.005, discount $\gamma = 0.99$, GAE $\lambda = 0.95$, and max gradient norm 1.0; advantages use GAE and are standardized. A low-frequency variant scales rollout length with the control period and exponentiates γ and λ to keep the effective per-second discount unchanged. All experiments are trained on 2 NVIDIA GeForce RTX 4090 48 GB GPUs with 30,000 iterations.

Goal Masking. During vocabulary-directed distillation, the student is never trained on the full tracking goal but instead on *masked* variants of it, inspired by the partially constrained controller in MaskedMimic. Let $g_t^{\text{track}} \in \mathbb{R}^{D_g}$ denote the teacher tracking target at time t. We introduce a binary visibility vector $m_t \in \{0,1\}^{D_g}$ and define

$$\tilde{g}_t = m_t \odot g_t^{\text{track}}, \qquad \mathcal{M}(g_t^{\text{track}}) = [\,\tilde{g}_t,\, m_t\,], \tag{15}$$

where $\odot$ is elementwise multiplication and $[\,\cdot,\cdot\,]$ denotes concatenation. At each step, every scalar dimension d of the tracking goal is independently marked as visible or hidden by sampling

$$m_{t,d} \sim \text{Bernoulli}(p_{\text{vis}}), \tag{16}$$

where p_{vis} follows a monotonically decreasing training schedule. It is initialized to a nonzero value to provide useful tracking guidance during early optimization and is gradually annealed toward zero as training progresses. The resulting mask m_t determines which entries of g_t^{track} are revealed to the student (via $\tilde{g}_t$) and which tokens the Transformer prior is allowed to attend to (via its attention mask). To avoid frame-by-frame "flickering" of supervision and to encourage the policy to handle genuinely under-specified goals, a sampled mask pattern is reused for a short temporal window before being resampled, so that the same subset of dimensions remains consistently observed or unobserved over several consecutive steps.

This structured random masking acts as goal-space domain randomization: large p_{vis} recovers an almost fully constrained imitation regime, while small p_{vis} forces the prior to rely more heavily on motion tokens and self-observation. At *training* time we always sample m_t as above, whereas at *inference* time we disable the tracking-goal stream by setting $m_t = \mathbf{0}$ (i.e., no tracking goal is provided), so the deployed controller operates purely from motion vocabulary and proprioceptive inputs. Training over this spectrum of mask patterns encourages the controller to treat the tracking goal as an optional noisy hint, leading to a student that remains stable when explicit tracking targets are absent or partially missing at deployment.

Vocabulary Controller Training. We implement the vocabulary-directed student as a conditional VAE whose components are: an encoder $\mathcal{E}$, a Transformer prior ρ, and an action decoder $\mathcal{D}$. At each step, the input is assembled from (i) masked tracking-goal poses (ii) 512-d motion vocabulary embedding, and (iii) a self-observation token. Each stream passes through normalized MLPs (clamp

Table 5: Quantitative analysis of Motion CoT quality. We compare text-only LLM augmentation, raw VLM-generated CoT, and our iterative VLM-judge pipeline.

Method	Hallucination Rate↓	R-Precision↑	MM-Dist↓
LLM (text-only input)	0.23	0.290	5.719
Raw VLM	0.13	0.358	5.315
Iterative VLM Judge (Ours)	**0.02**	**0.447**	**4.911**

$= 5$) to produce 512-d tokens; visibility masks are mapped to the attention mask. The prior is a 4-layer, 4-head Transformer (feed-forward 1024, dropout 0.1) that parameterizes a Gaussian over the latent; the encoder provides a residual refinement MLP to this prior. The decoder is an MLP with layers 1024–1024–512 (ReLU, $\tanh$ head) that outputs normalized joint targets conditioned on the latent and current self-observation.

Training follows a teacher–student data-aggregation scheme: the student acts with masking enabled, a privileged tracking teacher supplies action labels, and we optimize an action reconstruction objective with a KL regularizer (annealed from 10^{-4} to 10^{-2} between epochs 3000 and 6000). We use 1024 parallel environments, 32-step rollouts, batch size 4096, 6 mini-epochs, Adam with learning rate 2×10^{-5}, and gradient clipping at 50.0. The target pose is visible with probability 0.1, and the vocab embedding is visible with probability 1.0 when present. At inference, latent noise is set to zero for deterministic control. Training and inference are performed on two NVIDIA GeForce RTX 4090 GPUs.

Sim-to-Real State Estimation. Following prior works (Flayols et al., 2017; Liao et al., 2023; 2025) , we estimate root linear velocity $\dot{\mathbf{p}}^{\text{root}}$ by combining a momentum observer with an Extended Kalman Filter over base pose, velocity, and Inertial Measurement Unit biases. This filtering ensures that both teacher and student policies operate on physically plausible proprioception, closing the sim-to-real gap.

A.1.3 Large Language–Action Model

Human Motion Data Augmentation. Previous work (Ouyang et al., 2025) has highlighted that the sparsity and abstractness of text annotations in the AMASS dataset limit unified modeling of motion and language. Designing denser, decomposable, and more specific annotations can significantly enhance motion understanding. Motivated by the text augmentation approach of Ouyang et al. (2025), which employs LLMs to generate densified textual descriptions from text alone, we leverage the multimodal large model Qwen2.5-VL (Bai et al., 2025) to jointly process textual descriptions and rendered motion sequences. This enables us to obtain more accurate motion CoT annotations, since a single high-level abstract description may correspond to multiple plausible motions, many of which do not align with the actual motion instance.

However, raw VLM-generated CoTs may still suffer from *hallucinations*, such as describing actions that are absent from the rendered motion or producing invalid output formats. To improve annotation reliability, we further introduce an iterative VLM-judge verification pipeline. Specifically, after Qwen2.5-VL generates an initial motion CoT, we use GPT-4o to verify whether the generated decomposition is consistent with both the original text description and the rendered motion sequence, and CoTs with obvious hallucinations, semantic inconsistencies, or formatting errors are filtered out or regenerated. As shown in Tab. 5, compared with text-only LLM augmentation and raw VLM augmentation, our iterative VLM-judge pipeline substantially reduces hallucinations and yields better downstream motion–language alignment. These results indicate that the proposed verification procedure produces more reliable CoT supervision for language-conditioned motion generation.

Supervised Fine-Tuning. We conduct supervised fine-tuning on the Qwen2.5-3B-Instruct architecture using our augmented human motion dataset. The model is initialized with pretrained weights from the base Qwen2.5-3B-Instruct model, preserving its original transformer architecture with 36 layers, 16 attention heads, and a hidden dimension of 2048. The vocabulary is extended through our custom tokenizer that incorporates both text and motion tokens, with a total vocabulary size of 152,180 tokens.

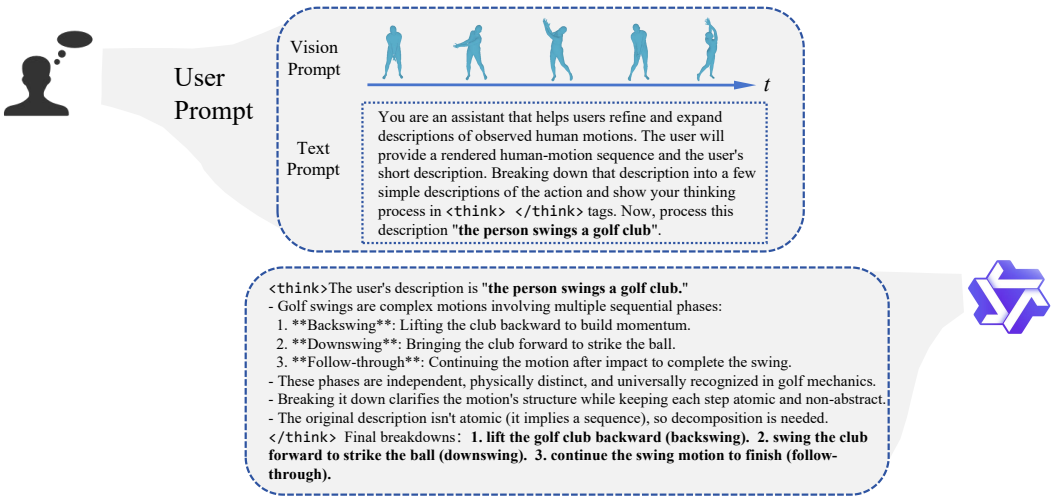


Figure 5: Illustration of dataset augmentation with Qwen2.5-VL (Bai et al., 2025). Given an AMASS (Guo et al., 2022) motion-text pair, we first render the motion sequence following Zhang et al. (2023), making it the vision prompt. Then, we obtain the text prompt by filling a predefined template using the raw textual annotation. Jointly, we query Qwen2.5-VL to get a decomposed, accurate, and semantic-enriched motion CoT.

For training data construction, we format each sample as a concatenated sequence of prefix tokens and target tokens. The prefix consists of the original text annotations formatted as instructional prompts, while the target comprises the motion CoT reasoning steps followed by the corresponding motion tokens.

Training is conducted on 8 NVIDIA H20 GPUs using AdamW optimization with a learning rate of 1×10^{-4} and weight decay of 0.01, following a cosine decay schedule with warmup. We employ mixed-precision training in bfloat16 and train for 30,000 steps with a fixed random seed to ensure reproducibility. The model is trained autoregressively with cross-entropy loss on continuous sequences, learning to generate motion reasoning and execution tokens conditioned on instructional prefixes without fixed context length constraints.

RL Fine-Tuning. We further fine-tune the model with reinforcement learning using the GRPO algorithm. Training is performed with batch size 64, consisting of 8 prompts per batch and 8 sampled responses per prompt, with gradient accumulation over micro-batches of 4. The maximum prompt and generation lengths are set to 512 and 1024, respectively. We use the memory-efficient AdamW optimizer with learning rate 1×10^{-5}, weight decay 0.01, and $(\beta_1, \beta_2) = (0.9, 0.999)$, along with gradient clipping at 1.0 and a cosine learning rate schedule decayed to 1×10^{-6}. The clipped objective adopts $\epsilon = 0.2$ and includes a KL regularization term with $\beta = 0.001$ against the SFT model as reference. The weighting coefficients in equation 11 are set as $\lambda_m = \lambda_t = 10$. We calculate $r_{\text{pos}} = \exp\left(-0.005\,\text{MPJPE}\right), r_{\text{acc}} = \exp\left(-0.05\,\text{E}_{\text{acc}}\right)$ in equation 12. The contrastive motion encoder ϕ_m and text encoder ϕ_t in equation 11 are trained following Guo et al. (2022) on a tracking-based retargeted AMASS dataset, where only successfully tracked sequences are retained. The training is also conducted on 8 NVIDIA H20 GPUs.

A.1.4 Baseline Implementations

In the following subsections, we summarize practical details of baseline implementations, evaluation metrics, and robot system setup used in Sec. 4. For a fair comparison with the baselines, we unify the evaluation metrics and protocols across all methods. The comparisons span representative paradigms (diffusion, VAE and RL), demonstrating the generality of our evaluation. To measure generation metrics, we adopt the motion and text encoders described in A.1.3. The implementation details for each baseline are outlined as follows:

MDM+Retarget. We employ MDM (Tevet et al., 2023), a representative diffusion-based motion generator, pretrained on the HumanML3D (Guo et al., 2022) dataset to generate human motions. Since MDM outputs joint positions rather than SMPL parameters, we apply an optimization to regress SMPL parameters from joint positions, using a learning rate of 10^{-1} for 100 iterations per generated sequence. The resulting SMPL sequences are then retargeted to the humanoid robot via a keypoint-based optimization proposed in H2O (He et al., 2024). As this baseline does not involve physical simulation, we report only generation metrics.

OmniH2O. OmniH2O (He et al., 2025b) extends MDM+Retarget by tracking generated kinematic humanoid motion sequences within a physics simulator using goal-conditioned reinforcement learning. Following the evaluation protocol of He et al. (2025b), we report physics-based tracking metrics computed over all sequences, rather than only the successfully tracked ones. After tracking, all trajectories are collected to compute generation metrics.

UH-1. UH-1 (Mao et al., 2025) is a humanoid motion generator based on an autoregressive transformer.We implement UH-1 on our physically retargeted humanoid motion dataset, which is constructed using Mink retargeting and tracking with the teacher controller described in Sec. 3.2. We first train the humanoid motion generator with a decoder-only transformer (Zhang et al., 2023), then leverage goal-conditioned RL to track these generated outputs. Generation metrics and physics metrics are reported following the same calculation paradigm as in the implementation of OmniH2O.

LangWBC. We reproduce LangWBC (Shao et al., 2025) by using a prior-free C-VAE student (no prior network) and a frozen CLIP text encoder (Radford et al., 2021b) to jointly encode language and proprioception and decode normalized joint targets. For policy distillation we use the same teacher controller in Sec. 3.2. Because LangWBC does not directly generate a reference tracking goal, we report only its success rate; MPJPE, E_{vel}, and E_{acc} cannot be calculated.

RLPF. We select RLPF (Yue et al., 2025) as an important baseline as it represents a newly introduced framework of leveraging RL to fine-tune a text-to-humanoid generator. For RLPF, we report the physics-based metrics of the RLPF-PHC variant, as its experimental setup and evaluation protocol are most comparable to ours and other baselines. We do not include generation metrics, since RLPF evaluates them in the human motion domain, whereas our evaluation is defined directly on humanoid motions.

UniAct-T2M. We retrain the UniAct (Jiang et al., 2025a) model using our proposed motion representation and unified motion vocabulary. We follow the same evaluation protocol described above for both generation and physics-based metrics.

A.1.5 Evaluation Metrics

We provide details of the metrics used for text-to-humanoid evaluation. Metrics are divided into two categories: generation-side metrics (Guo et al., 2022; Zhang et al., 2023), which measure semantic alignment and distributional fidelity, and physics-based tracking metrics (Luo et al., 2023), which assess physical executability in simulation.

Clarification on MPJPE. For physics-based tracking evaluation, we define the *reference* joint positions as the kinematic targets produced by decoding the generated motion tokens into the humanoid canonical motion representation, and we define the *predicted* joint positions as those obtained from the actual simulator rollout of the tracked humanoid. MPJPE is computed frame-wise between these two joint position sequences after temporal alignment, and then averaged over all frames and test samples.

Let $\mathcal{G} = \{(w_i^{(g)}, m_i^{(g)})\}_{i=1}^{N_g}$ denote the generated (text, motion) pairs, and $\mathcal{T} = \{(w_j^{(t)}, m_j^{(t)})\}_{j=1}^{N_t}$ denote the ground-truth test pairs. We denote the text encoder by $\phi_t(\cdot)$ and the motion encoder by $\phi_m(\cdot)$. Motion inputs are standardized before encoding using dataset mean and standard deviation:

$$\tilde{m} = \frac{m - \mu}{\sigma}, \tag{17}$$

where μ, σ are the precomputed dataset mean and std.

Generation Metrics.

- **Embeddings and pairwise distances.** For N paired samples, compute text and motion embeddings:
$$t_i = \phi_t(w_i), \qquad m_i = \phi_m(\tilde{m}_i), \quad i = 1, \ldots, N, \tag{18}$$
and form the Euclidean distance matrix $D \in \mathbb{R}^{N \times N}$ as
$$D_{ij} = \sqrt{\max\big(-2\, t_i^\top m_j + \|t_i\|_2^2 + \|m_j\|_2^2,\ 0\big)}. \tag{19}$$
- **R-Precision@k.** Measures whether the ground-truth motion for each text query is among the k nearest neighbors. Let $D_{i,(1)} \le D_{i,(2)} \le \cdots$ be the sorted distances in row i. Define
$$\mathbb{I}_{i,k} = \begin{cases} 1, & D_{i,i} \le D_{i,(k)}, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$
where $D_{i,i}$ is the distance to its paired motion. Then
$$\text{R-Precision@}k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{i,k}. \tag{21}$$
- **Matching score.** The average diagonal distance reflects text–motion alignment:
$$\text{MatchingScore} = \frac{1}{N} \sum_{i=1}^{N} D_{i,i}. \tag{22}$$
- **Diversity.** Measures intra-set variability of motion embeddings. Sample T unordered pairs $\{(p_r, q_r)\}_{r=1}^{T}$:
$$\text{Diversity} = \frac{1}{T} \sum_{r=1}^{T} \|m_{p_r} - m_{q_r}\|_2. \tag{23}$$
- **Fréchet Inception Distance (FID).** Compares generated vs. ground-truth motion embedding distributions. For mean and covariance
$$\mu = \frac{1}{N} \sum_{i=1}^{N} m_i, \qquad \Sigma = \frac{1}{N-1} \sum_{i=1}^{N} (m_i - \mu)(m_i - \mu)^\top, \tag{24}$$
the FID is
$$\text{FID}(\mathcal{T}, \mathcal{G}) = \|\mu_t - \mu_g\|_2^2 + \text{Tr}(\Sigma_t + \Sigma_g - 2(\Sigma_t \Sigma_g)^{1/2}). \tag{25}$$

Physics-Based Tracking Metrics.

- **Success rate (Succ.).** Fraction of sequences tracked without falling or excessive deviation:
$$\text{Succ.} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\Big[\max_t \|J_t^{(i)} - \hat{J}_t^{(i)}\|_2 < 0.5\ \text{m}\Big], \tag{26}$$
where $J_t^{(i)}$ and $\hat{J}_t^{(i)}$ denote reference and simulated joint positions.
- **Mean Per-Joint Position Error (MPJPE).** Average distance between predicted and reference joint positions:
$$\text{MPJPE} = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \|J_{t,j}^{(i)} - \hat{J}_{t,j}^{(i)}\|_2. \tag{27}$$
- **Velocity error (E_{vel}).** Discrepancy in per-joint velocities:
$$E_{\text{vel}} = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \|\dot{J}_{t,j}^{(i)} - \dot{\hat{J}}_{t,j}^{(i)}\|_2. \tag{28}$$
- **Acceleration error (E_{acc}).** Discrepancy in per-joint accelerations:
$$E_{\text{acc}} = \frac{1}{NTJ} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \|\ddot{J}_{t,j}^{(i)} - \ddot{\hat{J}}_{t,j}^{(i)}\|_2. \tag{29}$$

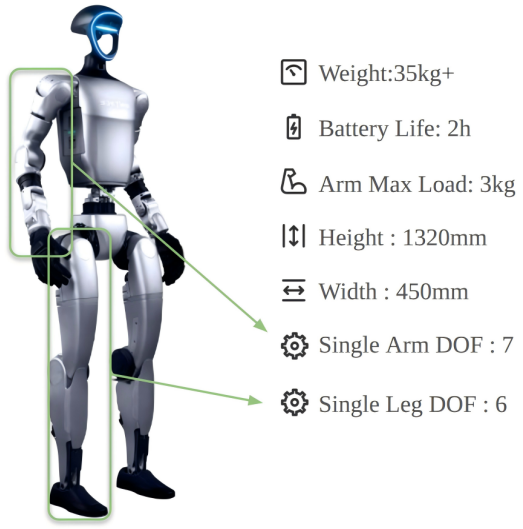


Figure 6: Unitree G1 Robot Hardware Details.



(a) HumanML3D annotations

(b) VLM-augmented CoTs

Figure 7: Word clouds of the original HumanML3D (Guo et al., 2022) text annotations (left) and our VLM-augmented motion CoT (right).

A.1.6 Robot System Setup

Our real humanoid hardware is built on the Unitree G1 humanoid platform (Robotics, 2024), as shown in Fig. 6. The G1 stands 1320 mm tall, weighs about 35 kg with battery, and provides 6 DOF per leg and 7 DOF per arm, with a maximum arm payload of 3 kg. It is powered by a 9000 mAh detachable battery that supports around 2 hours of operation. The knee joint could achieve up to 139 N·m torque with a joint movement space of 0–165°, complemented by other flexible joints including the waist (Z: $\pm 155°$, X: $\pm 45°$, Y: $\pm 30°$), the hip (P: $\pm 154°$, R: $-30°$ to $+170°$, Y: $\pm 158°$), and the wrist (P: $\pm 92.5°$, Y: $\pm 92.5°$), ensuring both stability and dexterity in whole-body control. We adopt the state-estimation and low-level deployment infrastructure of Liao et al. (2025) and Ze et al. (2025b), while the control commands are produced directly by our vocabulary-directed policy; no motion retargeting or reference-trajectory tracking is performed online.

A.2 Additional Experiments and Results

A.2.1 Dataset Curation

Fig. 5 illustrates our pipeline for generating motion CoT to augment the raw dataset using both visual and textual prompts. This section provides additional statistical details about the augmented dataset. Since our text-annotated AMASS subset originates from HumanML3D (Guo et al., 2022), we apply the chain-of-thought augmentation to the entire HumanML3D dataset. We employ `mink`

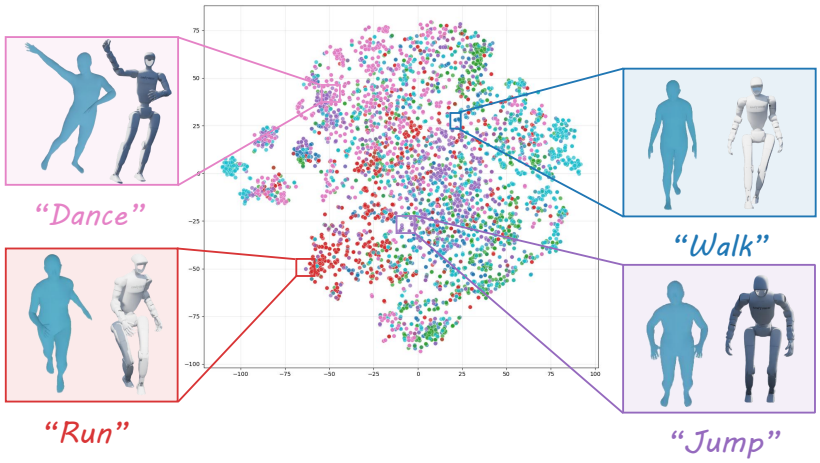


Figure 8: t-SNE visualization of different motion categories.

and smoothness filters for motion retargeting, resulting in an augmented dataset containing human-humanoid-text-CoT quadruplets.

After augmentation, our dataset exhibits substantial enrichment in textual descriptors. The vocabulary size expands from 5,371 to 11,896 distinct words, while the average textual length per sample increases from 12 to 163 words, reflecting the detailed nature of the motion CoT. Fig. 7 displays the 30 most frequent words in HumanML3D motion descriptions and our generated motion CoT annotations, respectively. The visualization indicates that the augmentation process enriches the vocabulary with high-level semantic concepts (e.g., *"posture", "locomotion", "static", "focuses"*). This enhancement not only increases the semantic diversity of the textual annotations but also facilitates our model's understanding of high-level instructions.

A.2.2 Analysis of Unified Motion Vocabulary

In this section, we present more direct and visually interpretable results to showcase the efficacy of our proposed unified vocabulary from answering two questions: (1) *Does the learned latent space have a good semantic structure?* and (2) *Can the learned latent space be semantically aligned between human and humanoid robot?*

To answer these two questions, we begin by defining six motion categories (walk, turn, run, jump (hop), sit, dance). Motion sequences are grouped according to these categorical labels derived from original motion annotations. For each category, we quantize the motion sequences using the pre-trained unified motion vocabulary. Each motion sequence is represented as a 512-dimensional latent vector by averaging along the time axis. To ensure balanced visualization across categories, we sample an equal number of instances from each class, determined by the size of the smallest category.

For the first question, we use t-SNE to demonstrate the discriminative capability of our latent space across different motion categories. The t-SNE results in Fig. 8 show that while some categories form distinct clusters, others exhibit some overlap. This can be attributed to the nature of the motion dataset: many sequences inherently encompass multiple modes of movement, meaning they naturally belong to several of our defined categories simultaneously. Consequently, the latent representations of these multi-modal sequences appear in overlapping regions of the t-SNE plot. For instance, "walk" (blue) is a common motion that frequently co-occurs with other motions in annotations, leading to its widespread distribution across the plot. In contrast, "dance" (red) is a more distinctive and less frequently combined motion, resulting in a more concentrated cluster. This overlap reflects the rich compositional nature of human motion rather than a limitation of the latent space. For the second question, within each category, we decode the same motion code into human and humanoid motion respectively by modal-specific decoders to visualize cross-embodiment alignment capability.

A.2.3 Additional Evaluations

More Details About Free-form Instructions. The LLM-augmented set contains 100 prompts. We first manually designed 10 examples spanning distinct whole-body activities, then used them as references to prompt DeepSeek (Shao et al., 2024) for 100 additional instructions in a consistent style. Tab. 6 shows representative prompts from this set. Each prompt was evaluated over five simulation trials. The human-written set contains 100 prompts sampled from the human-annotated subset of SnapMoGen (Hwang et al., 2026); these prompts were neither generated nor filtered by an LLM or VLM. This second source tests whether the results extend beyond the manually seeded, LLM-augmented prompt style.

Table 6: Representative prompts from the LLM-augmented free-form instruction set.

Free-form Instructions	
1. A conductor leads an orchestra with sweeping arm movements.	6. A teacher writes on an invisible blackboard, then turns to address the class.
2. A police officer is directing traffic at an intersection.	7. Twist your waist to the left and then to the right, like warming up.
3. Walk with a confident and happy strut.	8. A boxer punched forward continuously.
4. A dancer performs a series of ballet turns and graceful arm movements.	9. A martial artist practices a slow, flowing sequence of Tai Chi moves.
5. A warrior gracefully performs a ritualized sword dance.	10. A rock star plays an epic air guitar solo with high energy and passion.

We use the success-rate (Succ.) and motion–text distance (MM-Dist) definitions of Sec. A.1.5 on both sets. Succ. measures physical execution, whereas MM-Dist measures whether the executed motion matches the prompt; reporting both helps expose physically feasible but semantically generic motions. LangWBC* uses our reproduced implementation described in Appendix A.1.4.

On the LLM-augmented set (left panel of Tab. 3), Humanoid-LLA reaches 74% success, 8 percentage points above the strongest baseline, FRoM-W1 (66%). Its MM-Dist of 6.12 is 2.46 lower than the best baseline value of 8.58, also from FRoM-W1. Compared with our no-CoT ablation, success increases from 68% to 74% and MM-Dist falls from 8.77 to 6.12, consistent with better motion–language alignment from motion CoT.

On the human-written set (right panel of Tab. 3), Humanoid-LLA obtains 78% success versus 66% for the strongest baseline, TextOp, and reduces MM-Dist from TextOp's 9.04 to 6.44. Relative to our no-CoT ablation, success increases by 13 percentage points and MM-Dist falls from 9.21 to 6.44. The gains on prompts sourced from human annotations complement those on the LLM-augmented set.

Evaluation with an Independent Text–Motion Encoder. To address concerns about reward hacking and metric overfitting, we recompute FID, R-Precision, and MM-Dist with a separately trained TMR-style evaluator based on pretrained T5-XXL features. This evaluator is independent of the encoder used for the training reward. Under the independent evaluator, Humanoid-LLA retains clear generation-quality gains over LangWBC and UniAct-T2M across all three metrics. This shows that the observed improvement is not an artifact of exploiting the training reward encoder or overfitting to the original evaluation metrics.

Table 7: Generation metrics recomputed with a separately trained TMR-style evaluator.

Method	FID↓	R-Prec.↑	MM-Dist↓
LangWBC (Shao et al., 2025)	8.43	0.34	6.79
UniAct-T2M (Jiang et al., 2025a)	9.11	0.30	7.18
Humanoid-LLA (Ours)	**5.72**	**0.43**	**5.60**

Comparison with more Text-Conditioned Baselines. We further compare with TextOp and FRoM-W1 under the matched retraining and evaluation setup described in Sec. A.1.4. Tab. 8 reports semantic generation and simulated physical execution metrics on the same prompts and G1 representation.

Table 8: Comparison with retrained TextOp and FRoM-W1 under the same HumanML3D split, G1 representation, prompts and simulation-based metric protocol.

Method	FID↓	R-Prec.↑	MM-Dist↓	Succ.↑	MPJPE↓	E_{vel} ↓	E_{acc} ↓
TextOp (Xie et al., 2026)	8.191	0.308	5.827	78.0%	61.32	9.45	8.96
FRoM-W1 (Li et al., 2026)	4.837	0.378	5.302	75.3%	63.09	10.70	10.14
Humanoid-LLA (Ours)	**2.626**	**0.447**	**4.911**	**87.6%**	**56.43**	**8.92**	**8.74**

Under this protocol, Humanoid-LLA has the best reported values in all seven columns of Tab. 8. Its simulated tracking success exceeds TextOp by 9.6 percentage points and FRoM-W1 by 12.3 points. Together with the component ablations in Tab. 2, these results support the benefit of combining semantic supervision with a token-conditioned controller and physical-feedback fine-tuning.

Inference Latency. We further evaluate the inference latency of Humanoid-LLA to assess its practicality for instruction-triggered humanoid control. All measurements are conducted on a single NVIDIA H20 GPU using the vLLM framework. We consider two representative settings: (i) in-distribution instructions from HumanML3D and (ii) out-of-distribution free-form instructions. The reported latency adds the high-level LLA generation time, which includes instruction understanding and motion-token generation, to one low-level vocab-directed controller inference step. It therefore measures the delay from a command to the first action, rather than the duration of the full motion execution.

As shown in Tab. 9, Humanoid-LLA has a first-action latency of approximately 1.8 seconds per instruction in both settings, while a controller inference step takes about 7.8 ms. These results suggest that the current system is suitable for instruction-triggered whole-body behaviors, where the language model generates a motion plan at the command level and the controller executes it in real time. Further reducing the high-level generation latency, for example by adopting lighter language backbones or more efficient decoding strategies, remains a promising direction for future real-time interactive deployment.

Sim-to-Real Humanoid Deployment. To evaluate the model's capability in Sim2Sim transfer and Sim2Real deployment, we conduct additional experiments on seven classical tasks across three physical settings: Sim-Only, Sim2Sim, and Sim2Real, as summarized in Tab. 10.

For each task, we collect relevant motion sequences by retrieving from a motion annotation database. Specifically, we query the database using task-specific keywords (e.g., "walk") and label all motion sequences whose annotations contain the keyword as belonging to that task. For the first six tasks, we randomly select five motion samples from the retrieval results. For the final task, "Kung fu", only two samples are selected due to limited data availability.

In the Sim-Only setting, we perform 2,048 trials in IsaacLab, while the Sim2Sim transfer is evaluated with 256 trials in MuJoCo. Considering the practical costs of real-world deployment, we conduct 10 trials per task for the Sim2Real evaluation. For real-world trials, a trial is considered successful only if the robot does not fall and does not exhibit significant deviation from the intended motion trajectory. The success rates and average motion lengths for all tasks are reported in Tab. 10.

The results reveal a consistent performance degradation from Sim-Only to Sim2Real, with simpler motions like "Walk" maintaining high success rates across domains. In contrast, dynamic motions such as "Dance" and "Kung fu" show significant Sim2Real gaps, particularly for longer durations where error accumulation amplifies. This highlights the challenge of preserving complex motion fidelity in real-world transfer.

Cross-Embodiment Transfer. To evaluate transfer of the shared motion vocabulary, we conduct experiments on the Booster T1 platform, with qualitative examples in Fig. 3. We augment vocabulary construction with T1 trajectories and train a T1-specific vocab-directed controller following the same distillation pipeline used for Unitree G1. The SFT-pretrained LLA model is reused without

Table 9: Inference latency of Humanoid-LLA. Gen. Avg denotes LLM generation time per instruction, Ctrl. Avg denotes controller inference time per control step, and First-action Avg denotes their sum for the first action.

Setting	Gen. Avg (s)	Ctrl. Avg (ms)	First-action Avg (s)
HumanML3D	1.850	7.761	1.858
Free-form instructions	1.773	7.825	1.781

Table 10: Success rate on Sim2Real text-to-humanoid control.

Humanoid-LLA				
Tasks	**Length (s)**	**Sim-Only**	**Sim2Sim**	**Sim2Real**
Walk	6.8	100%	100%	100%
Turn	7.6	100%	97.3%	90%
Run	5.8	98.8%	95.3%	90%
Jump	5.2	92.3%	89.4%	60%
Kick	5.5	85.9%	82.0%	70%
Dance	9.1	78.3%	68.3%	40%
Kung fu	10.3	81.0%	74.2%	50%

retraining or fine-tuning on T1-specific language–motion pairs; its generated tokens are executed by the T1-specific controller.

We further compare this cross-embodiment configuration with a T1-specific retraining baseline under the same evaluation protocol. Tab. 11 reports physical success and tracking error together with motion–language alignment on Booster T1.

Table 11: Quantitative comparison on Booster T1 under the same evaluation protocol. Higher success and lower MPJPE and MM-Dist are better.

Method	Succ.↑	MPJPE↓	MM-Dist↓
T1-specific retraining	48%	102.05	16.32
Cross-embodiment (Ours)	**56%**	**93.88**	**12.75**

The cross-embodiment configuration improves success by 8 percentage points and lowers MPJPE from 102.05 to 93.88 and MM-Dist from 16.32 to 12.75. These results support the modularity of the shared human–robot vocabulary: the high-level language-to-token model transfers across embodiments, while each robot receives its own physical adaptation. This is transfer with T1 trajectories and a T1-specific controller, rather than zero-shot physical execution on a new robot.

A.2.4 Additional Ablations

Unified Tokenizer. To demonstrate the effectiveness of our implicit-partitioning tokenizer for fine-grained joint quantization of human and humanoid motion, we compare against T2M-GPT (Zhang et al., 2023), a representative baseline in motion quantization. We further ablate the number of sub-codebooks and the number of entries per sub-codebook, and we evaluate the effect of omitting the cross-reconstruction loss $\mathcal{L}_{\text{cross}}$. The human, humanoid, and human-to-humanoid motion reconstruction evaluations all follow the implementation of Zhang et al. (2023). We show these experimental results in Tables 12, 13, and 14.

Results show that increasing the number of sub-codebooks and enlarging per-codebook capacity reduce quantization error. Compared with single-codebook quantization, implicit partitioning produces a more fine-grained discrete latent space under the same total token budget. As reported in Table 14, removing the cross-reconstruction term $\mathcal{L}_{\text{cross}}$ in equation 1 substantially degrades the human-to-humanoid reconstruction metric, demonstrating that the cross-modal objective is essential for assigning the same discrete token to semantically equivalent motion primitives across embodi-

ments. For the full language-conditioned pipeline, we use 8 sub-codebooks with 64 entries each; the end-to-end capacity ablation in Tab. 15 explains this choice beyond reconstruction quality alone.

Table 12: Human motion reconstruction. N is the number of sub-codebooks, K the entries per book, and $\rightarrow$ means closer to the dataset distribution.

Methods	FID↓	TOP-1↑	TOP-2↑	TOP-3↑	Diversity→	MM-Dist↓
T2M-GPT	$0.112^{\pm.001}$	$0.500^{\pm.003}$	$0.692^{\pm.002}$	$0.789^{\pm.002}$	$9.723^{\pm.066}$	$3.056^{\pm.009}$
Ours (N=4, K=64)	$0.077^{\pm.000}$	$0.507^{\pm.003}$	$0.699^{\pm.003}$	$0.793^{\pm.003}$	$9.645^{\pm.098}$	$3.005^{\pm.008}$
Ours (N=8, K=32)	$0.032^{\pm.000}$	$0.509^{\pm.003}$	$0.702^{\pm.002}$	$0.796^{\pm.002}$	$9.563^{\pm.063}$	$2.986^{\pm.008}$
Ours (N=8, K=128)	$0.018^{\pm.000}$	$0.509^{\pm.002}$	$0.702^{\pm.002}$	$0.796^{\pm.002}$	$9.579^{\pm.084}$	$2.982^{\pm.009}$
Ours (N=16, K=64)	$0.005^{\pm.000}$	$0.510^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.003}$	$9.434^{\pm.069}$	$2.968^{\pm.009}$
Ours w/o $\mathcal{L}_{\text{cross}}$	$0.041^{\pm.000}$	$0.508^{\pm.003}$	$0.700^{\pm.002}$	$0.794^{\pm.002}$	$9.488^{\pm.072}$	$2.987^{\pm.009}$
Ours (N=8, K=64)	$0.021^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.002}$	$0.796^{\pm.002}$	$9.555^{\pm.056}$	$2.978^{\pm.009}$

Table 13: Humanoid motion reconstruction; notation and metric directions follow Tab. 12.

Methods	FID↓	TOP-1↑	TOP-2↑	TOP-3↑	Diversity→	MM-Dist↓
T2M-GPT	$0.183^{\pm.002}$	$0.475^{\pm.003}$	$0.661^{\pm.002}$	$0.758^{\pm.002}$	$10.804^{\pm.093}$	$3.425^{\pm.008}$
Ours (N=4, K=64)	$0.082^{\pm.001}$	$0.484^{\pm.003}$	$0.673^{\pm.002}$	$0.771^{\pm.002}$	$10.805^{\pm.096}$	$3.338^{\pm.008}$
Ours (N=8, K=32)	$0.037^{\pm.000}$	$0.491^{\pm.003}$	$0.679^{\pm.003}$	$0.776^{\pm.002}$	$10.577^{\pm.003}$	$3.307^{\pm.008}$
Ours (N=8, K=128)	$0.016^{\pm.000}$	$0.492^{\pm.003}$	$0.680^{\pm.003}$	$0.777^{\pm.002}$	$10.537^{\pm.094}$	$3.291^{\pm.008}$
Ours (N=16, K=64)	$0.006^{\pm.000}$	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.003}$	$10.653^{\pm.080}$	$3.288^{\pm.008}$
Ours w/o $\mathcal{L}_{\text{cross}}$	$0.011^{\pm.000}$	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.003}$	$10.631^{\pm.093}$	$3.285^{\pm.008}$
Ours (N=8, K=64)	$0.023^{\pm.000}$	$0.490^{\pm.003}$	$0.678^{\pm.003}$	$0.776^{\pm.002}$	$10.671^{\pm.089}$	$3.301^{\pm.008}$

Table 14: Human-to-humanoid motion reconstruction; notation and metric directions follow Tab. 12.

Methods	FID↓	TOP-1↑	TOP-2↑	TOP-3↑	Diversity→	MM-Dist↓
T2M-GPT	$0.381^{\pm.001}$	$0.460^{\pm.002}$	$0.642^{\pm.002}$	$0.741^{\pm.002}$	$10.689^{\pm.082}$	$3.540^{\pm.007}$
Ours (N=4, K=64)	$0.227^{\pm.002}$	$0.468^{\pm.003}$	$0.655^{\pm.003}$	$0.754^{\pm.003}$	$10.687^{\pm.083}$	$3.445^{\pm.008}$
Ours (N=8, K=32)	$0.182^{\pm.002}$	$0.470^{\pm.003}$	$0.657^{\pm.003}$	$0.755^{\pm.002}$	$10.777^{\pm.109}$	$3.450^{\pm.007}$
Ours (N=8, K=128)	$0.107^{\pm.001}$	$0.476^{\pm.002}$	$0.665^{\pm.002}$	$0.764^{\pm.002}$	$10.704^{\pm.099}$	$3.387^{\pm.007}$
Ours (N=16, K=64)	$0.084^{\pm.001}$	$0.480^{\pm.003}$	$0.668^{\pm.003}$	$0.766^{\pm.002}$	$10.701^{\pm.060}$	$3.377^{\pm.007}$
Ours w/o $\mathcal{L}_{\text{cross}}$	$25.044^{\pm.028}$	$0.074^{\pm.001}$	$0.138^{\pm.002}$	$0.192^{\pm.002}$	$6.230^{\pm.048}$	$8.192^{\pm.009}$
Ours (N=8, K=64)	$0.153^{\pm.002}$	$0.477^{\pm.002}$	$0.665^{\pm.002}$	$0.762^{\pm.002}$	$10.736^{\pm.102}$	$3.396^{\pm.006}$

Codebook Capacity and End-to-End Control. With eight sub-codebooks, even 64 entries per book allow a nominal 64^8 combinations of motion tokens. The reconstruction results above show that increasing per-book capacity can improve motion reconstruction, but this need not improve the complete language-conditioned system. We therefore compare 64, 1024, and 2048 entries per sub-codebook while keeping the number of sub-codebooks at eight.

Table 15: End-to-end codebook-capacity ablation with eight sub-codebooks (K entries per book).

K	Succ.↑	MPJPE↓	$E_{\text{vel}}\downarrow$	$E_{\text{acc}}\downarrow$
1024	84.8%	57.92	9.10	**7.27**
2048	82.1%	59.55	9.03	7.82
64 (Ours)	**87.6%**	**56.43**	**8.92**	8.74

Compared with $K = 64$, larger codebooks reduce success rate and increase MPJPE, although $K = 1024$ achieves the lowest acceleration error. These results favor $(8, 64)$ for overall downstream control performance. Sparser observations per token may make autoregressive prediction and token-conditioned control harder.

Vocabulary-Directed Action Distillation. Table 16 studies three key components of the vocabulary-directed student: a VAE latent, a Transformer prior, and random mask training. Replacing the **VAE** to MLP causes the largest drop in executability and accuracy: success falls from $95.2/86.1\%$ (train/test) to $93.8/84.6\%$, while MPJPE degrades markedly ($39.86 \rightarrow 59.24$ train; $49.69 \rightarrow 68.57$ test), and $E_{\text{acc}}/E_{\text{vel}}$ increase (train: $6.88/6.13 \rightarrow 8.31/7.37$; test: $9.23/8.18 \rightarrow 11.73/10.84$). Dropping the **prior** also hurts but less severely: success $94.1/85.3\%$ and MPJPE $48.72/59.91$, indicating the prior supplies helpful dynamics regularization under token guidance. Eliminating the **mask** leads to the lowest test success (83.5%) and higher errors (MPJPE 61.83, $E_{\text{acc}} = 9.98$, $E_{\text{vel}} = 9.02$), suggesting that masking mitigates overfitting to dense teacher signals and encourages reliance on discrete tokens. Overall, the vocabulary-directed student approaches the teacher's success while accepting slightly higher pose errors. After integrating the student into the full RL pipeline, **Humanoid-LLA** surpasses the teacher in test success ($\mathbf{87.6}\%$ vs. 86.1%) and reduces E_{acc} from 9.23 to 8.92, while E_{vel} rises from 8.18 to 8.74.

Table 16: Ablations on Vocabulary-directed Action Distillation.

	HumanML3D-Train				HumanML3D-Test			
Methods	Succ. ↑	MPJPE ↓	E_{vel} ↓	E_{acc} ↓	Succ. ↑	MPJPE ↓	E_{vel} ↓	E_{acc} ↓
Ours Tracking Controller	95.2%	39.86	6.88	6.13	86.1%	49.69	9.23	8.18
Ours w/o VAE	93.8%	59.24	8.31	7.37	84.6%	68.57	11.73	10.84
Ours w/o prior	94.1%	48.72	8.19	7.28	85.3%	59.91	9.56	8.71
Ours w/o mask	92.9%	51.32	8.62	7.58	83.5%	61.83	9.98	9.02
Humanoid-LLA (Ours)	95.0%	46.84	8.04	6.86	87.6%	56.43	8.92	8.74

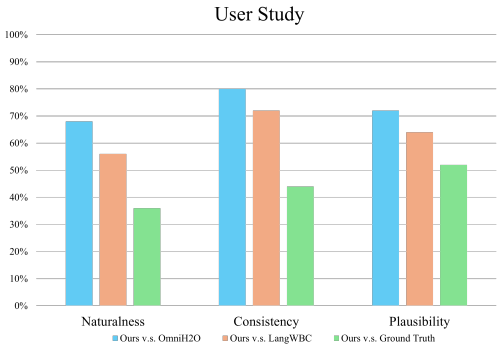
A.2.5 User Study

To evaluate the perceptual quality of the generated robot motions, we conduct a user study with 25 participants (N=25). For each trial, a natural-language instruction is shown together with two motion clips rendered side-by-side: one produced by our method and one produced by a reference method (OmniH2O (He et al., 2025b), LangWBC (Shao et al., 2025)), or the recorded ground-truth motion. The order of the two clips is randomized across trials to avoid ordering bias.

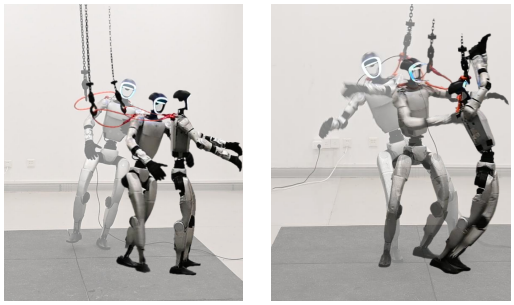
Participants are asked to answer three pairwise comparison questions for each motion pair, as summarized in Tab. 17. Specifically, they are asked (i) which motion looks more natural and human-like (Naturalness), (ii) which motion better matches the text description (Consistency), and (iii) which motion appears more physically plausible for the robot to execute (Plausibility). For each question, participants select exactly one of the two motions.

We aggregate all votes and report the preference rate for our method in Fig. 9a. Overall, our motions are preferred to OmniH2O (68%, 80%, 72%) and LangWBC (56%, 72%, 64%) in terms of naturalness, consistency, and plausibility, and even approach the ground-truth motions (36%, 44%, 52%) on these metrics.

These results show that Humanoid-LLA not only improves alignment between language and motion, but also produces robot behaviors that users perceive as more natural, more semantically consistent,



(a) User study results.



(b) Failure cases. Left: unstable single-leg support causes directional deviation. Right: excessive force in one direction leads to instability.

Figure 9: User study and failure case analysis. Participants perceive our method as better across Naturalness, Consistency, and Plausibility, while failure cases mainly arise from unstable single-leg support or excessive directional force.

and more physically plausible than those generated by strong baselines, and close to human-authored demonstrations.

Table 17: Question settings for user questionnaire. We evaluate the motion generated by different methods in three metrics: Motion Naturalness, Text Consistency, and Physical Plausibility.

Metrics	Question
Naturalness	Which motion looks more natural and human-like?
Consistency	Which motion better matches the text description?
Plausibility	Which motion is more physically plausible for the robot to execute?

A.2.6 Additional Text-to-Humanoid Visualizations

To further validate the effectiveness of our method, we provide more text-to-humanoid visualization in this section, including indoor and outdoor experiments on a physical humanoid robot (Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, and Fig. 15). The bottom of the real-world visualization figures shows the original text prompt, and the top shows the key portion extracted from motion CoT.

A.2.7 Supplementary Video

We provide a supplementary video that includes:

- Brief summary of our pipeline.
- Visualization of cross-embodiment alignments.
- Videos of real-world Text-to-Humanoid control.
- Visual comparison with LangWBC (Shao et al., 2025).
- Demonstration on real-world robustness.

A.3 Discussion and Limitations

A.3.1 Discussion of Design Choices

In this section, we position Humanoid-LLA within the broader landscape of text-to-humanoid control, and clarify both its connections to and differences from existing approaches, with a particular focus on how it leverages a unified motion vocabulary, a vocab-directed controller, and multimodally grounded

augmentation. Recent methods typically realize text-driven humanoid control by either (i) *directly regressing low-level humanoid poses or actions from text embeddings* (Mao et al., 2025; Shao et al., 2025), or (ii) *coupling a text-driven high-level human motion generator with a separately trained humanoid tracking controller* (He et al., 2025b; Ding et al., 2025; Yue et al., 2025). In contrast, Humanoid-LLA formulates text-to-humanoid control as **reasoning in a unified human–humanoid motion vocabulary**, where a language model predicts discrete motion tokens and a **vocab-directed controller**—distilled from a privileged tracking teacher and trained with **multimodally grounded augmentation** on paired language and visual/motion signals—realizes these tokens in a physically robust and semantically aligned way.

Role of Retargeting. A potential confusion is whether Humanoid-LLA is *retargeting-free*. Our claim is specifically about inference-time retargeting: unlike two-stage pipelines that must convert a newly generated human-space trajectory into humanoid joint targets for every test command, Humanoid-LLA does not require any such online kinematic projection. At deployment, the LLM outputs discrete motion tokens in a unified human–humanoid vocabulary, and a distilled vocab-directed controller executes these tokens directly with closed-loop proprioceptive feedback.

Retargeting is only used *offline* for two purposes: (i) to construct paired human–humanoid motion data that anchors token semantics across embodiments during vocabulary learning, and (ii) to train a privileged tracking teacher that validates token primitives under full-state supervision. This offline use does not contradict our goal; instead, it enables us to leverage abundant human motion data for language grounding while eliminating error accumulation and instability caused by *online* retargeting and tracking of noisy decoded references.

Vocabulary-Directed Policy Distillation. A central design choice in Humanoid-LLA is to distill a vocab-directed controller from a fully constrained tracking teacher, rather than driving the humanoid through a long chain of intermediate pose representations. Several existing text-to-humanoid pipelines (He et al., 2025b; Ding et al., 2025; Yue et al., 2025) first generate a *human-centric* motion representation, such as a 263-D HumanML3D-style feature sequence (Guo et al., 2022; Tevet et al., 2023; Jiang et al., 2024) on a canonical human skeleton, then apply a learned pose estimator to recover SMPL (Loper et al., 2015) pose parameters, followed by a retargeting algorithm (He et al., 2024; Araujo et al., 2025) to map SMPL poses to the humanoid, and finally run a tracking policy (Luo et al., 2023; Tessler et al., 2024; He et al., 2025b) to follow the retargeted trajectory. Each stage is trained with its own objective and inductive biases, so small mismatches between HumanML3D features, pose estimation, and retarget outputs can accumulate. In practice, the resulting humanoid trajectories can be brittle: even mild noise or local discontinuities in the estimated SMPL poses or in the retargeted motion may induce large errors for the low-level controller, leading to instability and low task success rates.

In contrast, Humanoid-LLA uses vocab-directed policy distillation to collapse this chain. A privileged teacher policy is trained to track offline-retargeted trajectories with full-state information and dense physics-based rewards, and its behavior is distilled into a student controller that takes only motion tokens and partial humanoid state as input and directly outputs joint-position targets for the low-level PD controller. At test time, the humanoid no longer depends on an online HumanML3D→SMPL→retarget pipeline; instead, the vocabulary-directed controller realizes each token as a short, dynamically feasible motion primitive that has been validated by the teacher. Importantly, our **goal in distillation** is *not* to surpass state-of-the-art retargeting algorithms in terms of raw tracking error, but to shield low-level control from the discontinuities and estimation noise introduced by pose estimation and retargeting, so that imperfections in the offline retargeted data do not translate into catastrophic failures during execution. This shift from tracking fragile per-frame pose targets to executing robust token-level primitives plays a key role in improving overall success rates in text-to-humanoid control.

Unified Motion Latent Space. Language and motion form an inherently many-to-many mapping: a single instruction can correspond to a diverse family of valid behaviors, and a single motion can be described by multiple semantically compatible utterances. Directly conditioning a behavior cloning or RL policy on continuous language embeddings (e.g., CLIP (Radford et al., 2021a) features or raw LLM token embeddings) effectively collapses this many-to-many relationship into a single network that must simultaneously encode high-level semantics and low-level control (Juravsky et al., 2022;

Shao et al., 2025). This tightly re-entangles the two modalities and makes it difficult to explicitly reason about the set of admissible motions for a given instruction.

We instead introduce a dedicated motion vocabulary that specializes in encoding motion primitives rather than language semantics. A dual-branch VQ-VAE first learns a discrete set of motion tokens from large-scale human motion data, and the same vocabulary is then refined with humanoid-specific feedback via the humanoid branch and the vocabulary-directed controller. This design allows us to amortize skill learning over abundant human data while still grounding the latent space in physically valid humanoid behaviors. Compared to text-embedding–conditioned policies, our GRPO-style fine-tuning also benefits from this discrete interface: each token deterministically decodes to a short motion primitive, so candidate token sequences can be rolled out, decoded into full trajectories, and evaluated with *explicit tracking rewards* in the humanoid dynamics space, instead of relying on indirect proxies in a continuous language embedding space. Finally, operating in a tokenized motion space helps avoid sacrificing diversity for control success. For a fixed instruction, the LLA naturally samples different token sequences across rollouts, exploring multiple valid realizations of the command, while the vocabulary-directed controller enforces local dynamic feasibility for each generated primitive.

Motion Chain-of-Thought Prompting. Another key difference from prior text-to-humanoid control frameworks lies in how Humanoid-LLA structures the reasoning process. Instead of mapping language directly to motion tokens, we optionally prepend a motion CoT—a short, structured sequence of intermediate subgoals or motion phrases—to the supervision target. This design mirrors chain-of-thought prompting in language reasoning, but grounds the intermediate steps in motion semantics (e.g., "turn left $\rightarrow$ walk forward $\rightarrow$ slow down and stop") rather than purely textual explanations.

Motion CoT is not a mandatory intermediate for every prompt. In our pipeline, CoT is invoked *only* when the instruction is abstract, long-horizon, or compositionally rich; for already atomic, low-level descriptions, the model preserves the original command without further decomposition. This design avoids unnecessary verbosity and prevents over-structuring simple motions.

Compared to prior approaches that mainly rely on *purely language-side* LLM augmentation (Ouyang et al., 2025)—e.g., generating additional descriptions or pseudo-instructions without constraining them with real motion trajectories, which can easily drift away from the underlying data—our augmentation is explicitly grounded in both language and visual/motion signals. The intermediate motion phrases are derived from, or validated against, the paired motion sequences, so the augmented supervision remains tightly coupled to what the human actually does, thereby strengthening text–motion alignment.

In principle, such intermediate motion rationales help the LLA decompose compositional instructions into shorter token segments that the vocabulary-directed controller can reliably execute. We view motion CoT as a first step toward more interpretable and debuggable text-to-humanoid policies, and future work could study how automatically generated versus human-annotated chains and different multimodal grounding strategies affect performance and robustness.

A.3.2 Limitations and Future Work

Failure Cases. Although our approach produces physically plausible and semantically rich motions in most scenarios, we observe a characteristic failure mode induced by the offline human-to-humanoid retargeting pipeline. Our current setup uses the Mink inverse-kinematics solver (Zakka, 2024) as instantiated in MaskedMimic (Tessler et al., 2024), which tends to retarget human demonstrations to humanoid motions with overly human-like, nearly straight-knee lower-limb postures. This in turn biases the shared latent space toward straight-knee lower-body semantics; when the model issues highly dynamic upper-body motions while implicitly preserving these straight-knee configurations (e.g., vigorous arm swings with limited knee flexion), the controller can produce unstable lower-body behaviors, including loss of contact consistency and visible foot skating, as illustrated in Fig. 9b. We view this as a limitation of the retargeting stage rather than the LLA model itself, and leave to future work the use of more robot-centric, dynamics-aware retargeting schemes—explicitly accounting for contact constraints, joint limits, and torque margins—to train the shared latent space and controller so as to reduce this bias and improve stability in challenging whole-body tasks.

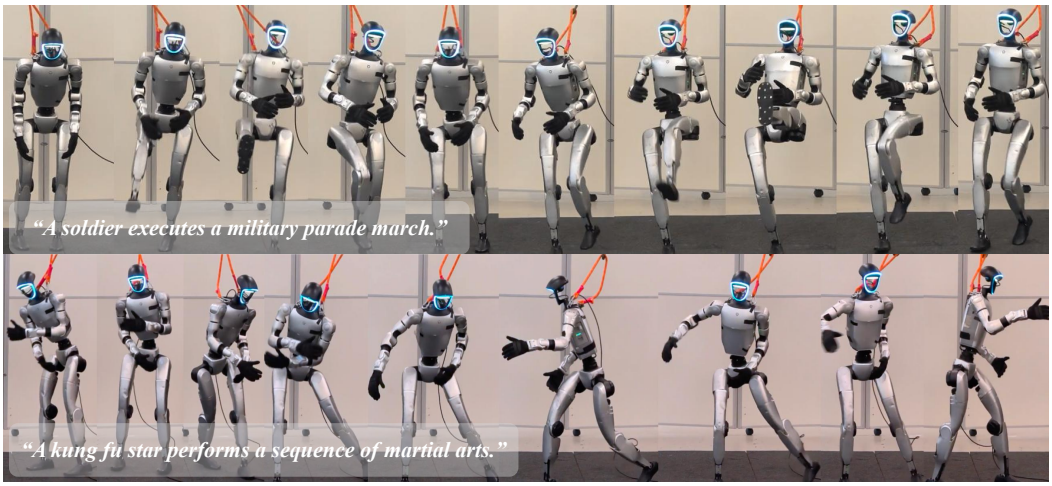


Figure 10: Results in the real world when prompted "A soldier executes a military parade march." and "A kung fu star performs a sequence of martial arts."

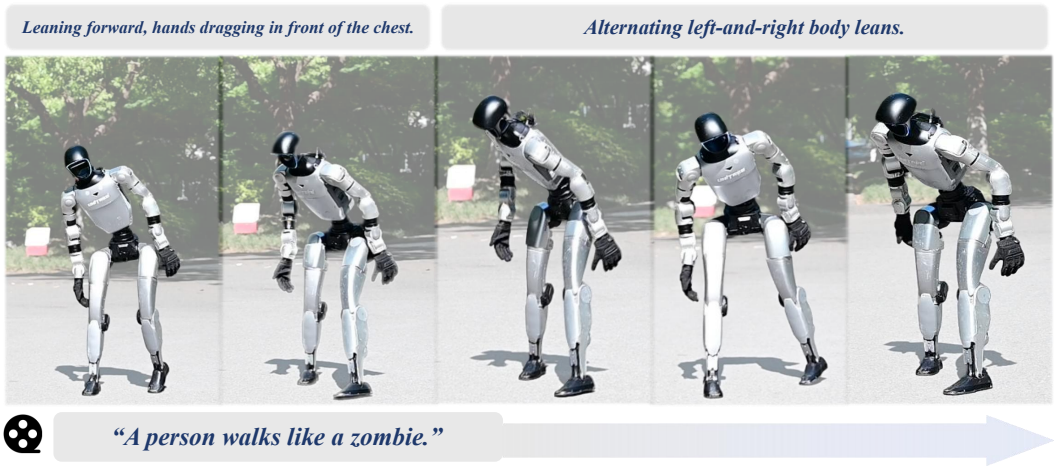


Figure 11: Results in the real world when prompted "A person walks like a zombie."

Future Work. In this work, we present a Large Language–Action Model for physical-fidelity, open-vocabulary humanoid control, mapping expressive natural language directly to executable humanoid actions. Through comparative experiments, we identify considerable areas where our model can be further improved, as outlined below:

- **Longer-Horizon Memory and Planning.** Our LLA reasons over token sequences within a modest temporal window, which limits plan consistency across complex tasks. A natural extension is to couple Humanoid-LLA with a hierarchical planner that maintains a persistent memory (e.g., key–value token cache or episode summaries) and proposes subgoals that the vocabulary-directed controller can realize, improving stability and global coherence over minutes-long activities.
- **Real-time Latency.** Our current pipeline prioritizes physically executable and high-level language understanding generation over interactive speed, and autoregressive decoding plus downstream tracking introduces non-trivial latency. Future work includes lightweight backbones, caching or streaming decoding, and distillation or retrieval-augmented short-horizon generation to reduce delay for real-time deployment.
- **Richer Multimodal Grounding.** We condition primarily on text (and optionally visual renders), while real deployments benefit from audio cues (speech prosody),



Figure 12: Results in the real world when prompted "A person leisurely plays golf."

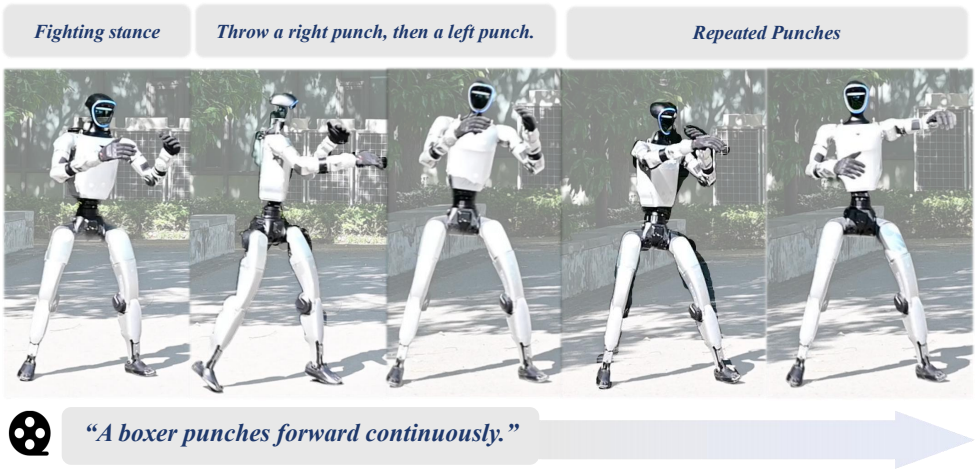


Figure 13: Results in the real world when prompted "A boxer punches forward continuously."

gaze/pose of humans, and tactile events. Extending the tokenizer with cross-modal slots (speech/vision/touch tokens) could align linguistic intent with environmental context, enabling disambiguation.

- **Personalization and Style Control.** Different users may prefer distinct motion styles or safety margins. Conditioning tokens on user embeddings (or few-shot style exemplars) can produce personalized motions while preserving safety. A style–safety Pareto controller could expose interpretable dials (conservativeness, speed, energy) without retraining.
- **Scaling Data and Benchmarks.** Our unified tokenizer aligns human and humanoid motions; scaling paired data with richer captions and hard negative text–motion pairs should improve semantic precision. We also advocate benchmarks that jointly score distributional quality and physics on robots, preventing degenerate solutions that optimize only one axis.

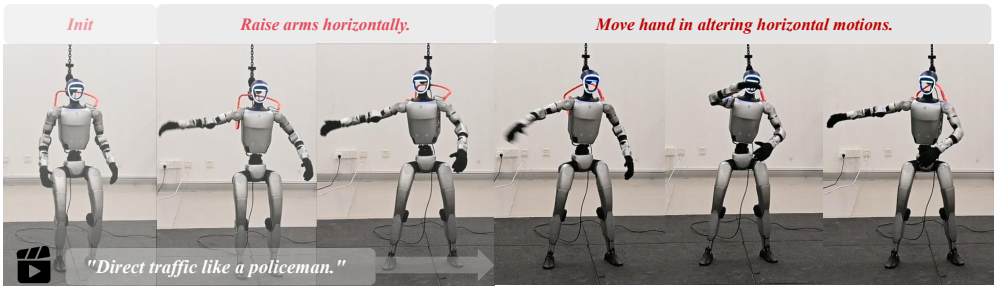


Figure 14: Results in the real world when prompted “Direct traffic like a policeman”.

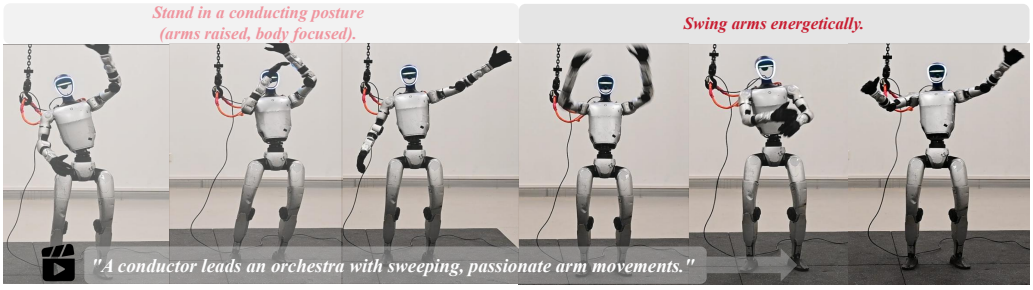


Figure 15: Results in the real world when prompted “A conductor leads an orchestra with sweeping, passionate arm movements”.