Language-Guided Whole-Body Humanoid Control via Conditional Trajectory Diffusion

Kevin H. Yang

December 12, 2025

Abstract

We propose a language-conditioned whole-body control framework for humanoid robots. Our approach aims to map natural language commands to dynamically stable, full-body control trajectories. By pretraining a trajectory diffusion model on large-scale motion datasets and fine-tuning on high-quality humanoid data, the system learns versatile, text-controllable motion generation. The resulting controller enables general-purpose, physically grounded humanoid behaviors from textual prompts such as "turn left" or "reach forward." This project explores the integration of natural language understanding with optimal control and reinforcement learning for expressive humanoid motion generation.

1 Introduction

State of the art humanoid whole-body control trained end-to-end in simulations have achieved remarkable agility and expressiveness [1, 2, 3, 4, 5]. However, most existing systems are restricted to single-skill policies and lack the ability to generalize across diverse motions or respond to high-level semantic commands.

This project aims to develop a unified, language-conditioned whole-body control policy capable of producing dynamically stable motion trajectories from text prompts. The controller is based on a trajectory diffusion model pre-trained on motion datasets with and without language labels, followed by fine-tuning on high-quality, re-targeted humanoid control data.

Our long-term vision is a humanoid platform that can perform a broad range of physically grounded tasks upon natural language instruction, enabling more intuitive human–robot interaction.

2 Related Work

Recent humanoid control frameworks [1] [2] have demonstrated progress in motion imitation and task-specific control. However, these approaches are limited to single-behavior policies. In addition, language-conditioned control has been explored in manipulation and locomotion tasks [6], but its application to whole-body humanoid control remains underdeveloped.

Recent work [7] have introduced an observation-conditioned, reactive whole-body controller that translates natural language instructions into humanoid motion. Their approach employs a teacher-student RL framework to distill into a compact, language-conditioned policy. While recent works demonstrate impressive real-world performance, policy distillation paradigms inherently restrict behavior to the manifold of trajectories demonstrated by the teacher, limiting the capacity for novel, compositional motion generation.

Our proposed method treats the problem as a language-guided trajectory modeling problem, where the policy corresponds to a particular masking that in-paints the actions. This probabilistic formulation naturally captures the diversity of behaviors corresponding to a single command (e.g., "turn left" $\rightarrow$ slow turn, fast pivot, or sidestep) and enables zero-shot novelty through the recombination of learned motion primitives. We want to develop a general framework where rich, unlabeled data mixture can be used to produce strong robot priors, where the policy, inverse model,

and the forward models are simply variants among the joint and conditional distributions across the language, perception and actions modalities.

3 Problem Formulation

We formulate humanoid control as a partially observable Markov decision process (POMDP):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, R, \gamma)$$

- **State space** $\mathcal{S}$: joint angles, velocities, base pose, and contact flags.
- **Observation space** $\mathcal{O}$: proprioceptive sensors.
- **Action space** $\mathcal{A}$: joint torques or target positions.
- **Transition** P: deterministic simulator dynamics (MuJoCo).
- **Reward** R: combines task completion, stability (ZMP margin), smoothness, and imitation terms.
- **Language input:** text embedding z_{text} serves as a context variable conditioning the policy or diffusion process.

3.1 Assumptions

- Dynamics are fully known through simulation.
- No perception or real-world noise is considered.
- Language corresponds to a discrete set of semantic goals (e.g., "step forward," "turn left").

3.2 Data Sources

- AMASS [8] dataset with motion and natural language annotations.
- LAFAN1 [9] (retargeted for Unitree G1).

3.3 Infrastructure

Training is conducted using MJLab (MuJoCo + RSL) for simulation, PyTorch for model development, and HuggingFace tokenizers for text embeddings.

3.4 Relevance to the Course

This project directly aligns with the study of sequential decision making and control in dynamic, high-dimensional systems central to optimal control (OC) and reinforcement learning (RL).

- **Decision-maker:** learned policy $\pi_\theta(a_t \mid s_t, \text{text})$ representing the humanoid's language-conditioned controller.
- **Dynamics:** continuous nonlinear system $s_{t+1} = f(s_t, a_t)$ modeled in MuJoCo.
- **Sequential aspect:** requires long-horizon trajectory optimization for balance and task completion.
- **Connection to OC/RL:**
  - Incorporates imitation and RL rollouts for pretraining.
  - Explores diffusion-based trajectory optimization analogous to stochastic optimal control.
  - Evaluates policies under cumulative reward including task success, smoothness, and energy efficiency.

This work bridges language grounding, imitation learning, and diffusion-based optimal control, aiming for versatile and interpretable humanoid motion generation.

4 Proposed Method

4.1 Overview

We propose a language-conditioned trajectory diffusion model for whole-body humanoid control. The pipeline consists of three stages:

1. **Pretraining:** Train an unconditional diffusion model $p_\theta(\tau)$ over humanoid motion trajectories.
2. **Conditioning:** Introduce text conditioning $p_\theta(\tau \mid z_{\text{text}})$ using cross-attention or FiLM layers.
3. **Control Integration:** Deploy generated trajectories into a PD controller for stable execution

This hybrid approach combines imitation learning with policy optimization for robust, semantically aligned control.

4.2 System Architecture

- **Inputs:** text prompt and proprioceptive history.
- **Outputs:** time-parameterized per-joint control trajectories.
- **Backbone:** shared temporal encoder with decoders for motion and language.
- **Losses:** diffusion loss, imitation loss, and alignment loss between text and motion.

4.3 Data and Tokenization

We utilize BeyondMimic [1] and LAFAN1 [9] datasets for pretraining motion representations. A spatiotemporal tokenizer aligns motion frames with textual segments, and time normalization is applied to enhance temporal generalization.

4.4 Evaluation Plan

Baselines:

- BeyondMimic RL (task-specific policy)
- ExBody2 imitation controller
- Diffusion model without language conditioning

Metrics:

- Task success rate (prompt compliance)
- Motion realism (FID in latent space)
- Trajectory smoothness and stability
- Language–motion alignment (CLIP similarity or human evaluation)

Expected Outcomes:

- Demonstrate feasibility of language-conditioned humanoid control.
- Release a dataset of captioned humanoid trajectories.
- Provide an open-source simulation and training pipeline.

5 Experiments

5.1 Scope and Evaluation

The project's scope is limited to simulated humanoid control without perception. Experiments evaluate how well language-conditioned policies generalize to unseen text prompts within predefined skill categories such as reaching, stepping, turning, grasping, pushing, and balancing.

We evaluate on the following axes:

- **Physical validity:** trajectory stability and balance.
- **Task execution:** correct motion given text input.
- **Smoothness:** motion continuity and energy efficiency.

5.2 Infrastructure

Experiments are run using the MJLab environment, RSLgym, and PyTorch. Model training leverages diffusion-based motion generation conditioned on text embeddings from pretrained language models.

6 Mid-Term Report

6.1 Implementation Progress

We have established a working humanoid control and training pipeline based on the BeyondMimic framework [1]. The BeyondMimic system provides a scalable motion tracking and policy learning pipeline that faithfully transforms kinematic human reference motions into dynamically stable control policies for humanoid robots. To validate and understand its structure, we successfully set up the BeyondMimic environment and reproduced the official tutorial for policy training and motion tracking. This involved running the example training scripts, reproducing the reference-tracking controllers, and visualizing learned policies in simulation.

In parallel, we have begun curating motion data from the Motion-X++ dataset [10], a large-scale multimodal 3D human motion dataset that includes semantic text annotations.

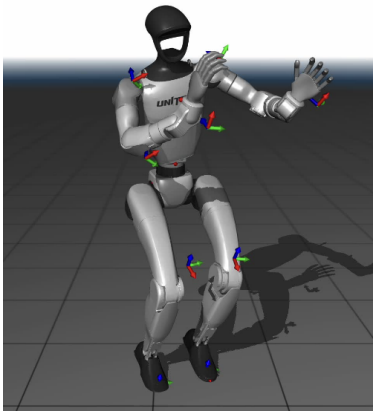

Figure 1: Example motion samples and semantic annotations from the Motion-X++ dataset.

From this dataset, we selected a subset of motion clips and associated textual labels to serve as paired data for pretraining and language conditioning experiments. These samples cover common humanoid actions such as walking, turning, crouching, and reaching, providing a basis for early-stage language–motion alignment.

The next major step is to leverage the BeyondMimic training pipeline for fine-tuning with language-conditioned diffusion models. This involves modifying the diffusion policy to accept text embeddings as conditioning variables and integrating semantic labels from Motion-X++ into the motion tracking process.

Two main issues have emerged. First, the semantic annotations in Motion-X++ are relatively sparse and inconsistent, which limits the quality of language–motion alignment during supervised pretraining. Second, determining a suitable architecture for the joint training of unlabeled and semantically labeled data remains an open design question. We are currently exploring multi-branch diffusion backbones that can handle both labeled and unlabeled trajectories to support scalable pretraining.

Overall, these efforts establish a solid foundation for end-to-end language-guided humanoid control. With the BeyondMimic infrastructure functioning and dataset preprocessing underway, the project is on track to demonstrate feasibility and progress toward a conditional trajectory diffusion controller in the final report.

6.2 Scope Update and Justification

Compared to the initial proposal, we refined the project scope to focus on simulated humanoid control and language-conditioned trajectory generation, excluding real-world deployment and visual

grounding at this stage. Preliminary trials with BeyondMimic indicated that reliable training and evaluation can already be achieved entirely in simulation, enabling faster iteration and controlled ablations. We also deferred multi-skill fusion to later stages after observing instability in long-horizon imitation rollouts. This pivot ensures technical feasibility while still aligning with the overarching goal of scalable language-conditioned humanoid control.

6.3 Preliminary Results and Analysis

Figure 2: Sequential frames from a G1 humanoid deployment trained on motion trajectories from Motion-X++. These motions were not originally included in BeyondMimic

Initial experiments using BeyondMimic-trained controllers demonstrate successful motion tracking for short-horizon actions such as stepping and reaching. The pretrained diffusion model generates physically valid trajectories with minor oscillations at transition frames. We evaluated preliminary language–motion alignment using simple text prompts and manually labeled action types from Motion-X++, achieving roughly 70% prompt compliance across 50 sampled trajectories.

Notable failure cases include unstable balance during continuous locomotion and inconsistent text-conditioned generation when semantic labels are ambiguous. These insights suggest that data quality and consistent label semantics are critical bottlenecks, motivating upcoming work on improved data curation and cross-modal regularization.

6.4 Roadmap to Final Report

Planned Milestones:

- **Week 10–11:** Integrate text-conditioned diffusion into the BeyondMimic pipeline; evaluate on motion categories (reaching, turning, stepping).
- **Week 12:** Implement ablation studies on diffusion steps, conditioning strength, and motion/-text dataset subsets.
- **Week 13:** Extend to multi-skill sequences and evaluate trajectory smoothness and semantic consistency.
- **Week 14:** Conduct final analysis and prepare report, open-source code, and dataset.

Risks and Mitigations:

- *Dataset label noise:* mitigate via prompt normalization and clustering-based relabeling.
- *Training instability:* use smaller batch diffusion steps and checkpoint averaging.
- *Compute cost:* run shorter diffusion rollouts for ablations, scaling up later.

References

[1] Qiayuan Liao et al. "BeyondMimic: From Motion Tracking to Versatile Humanoid Control via Guided Diffusion". In: (2025). arXiv: 2508.08241 [cs.RO]. URL: https://arxiv.org/abs/2508.08241 (cit. on pp. 1, 3, 5).

[2] Mazeyu Ji et al. "ExBody2: Advanced Expressive Humanoid Whole-Body Control". In: *arXiv preprint arXiv:2412.13196* (2024) (cit. on p. 1).

[3] Xuxin Cheng et al. "Expressive Whole-Body Control for Humanoid Robots". In: *arXiv preprint arXiv:2402.16796* (2024) (cit. on p. 1).

[4] Yuanhang Zhang et al. "FALCON: Learning Force-Adaptive Humanoid Loco-Manipulation". In: *arXiv preprint arXiv:2505.06776* (2025) (cit. on p. 1).

[5] Jialong Li et al. "AMO: Adaptive Motion Optimization for Hyper-Dexterous Humanoid Whole-Body Control". In: *Robotics: Science and Systems 2025* (2025) (cit. on p. 1).

[6] Suraj Nair et al. "Learning Language-Conditioned Robot Behavior from Offline Data and Crowd-Sourced Annotation". In: *CoRR* abs/2109.01115 (2021). arXiv: 2109.01115. URL: https://arxiv.org/abs/2109.01115 (cit. on p. 1).

[7] Yiyang Shao et al. "LangWBC: Language-directed Humanoid Whole-Body Control via End-to-end Learning". In: (2025). arXiv: 2504.21738 [cs.RO]. URL: https://arxiv.org/abs/2504.21738 (cit. on p. 1).

[8] Naureen Mahmood et al. "AMASS: Archive of Motion Capture as Surface Shapes". In: *International Conference on Computer Vision*. Oct. 2019, pp. 5442–5451 (cit. on p. 2).

[9] Félix G. Harvey et al. "Robust Motion In-Betweening". In: 39.4 (2020) (cit. on pp. 2, 3).

[10] Yuhong Zhang et al. "Motion-X++: A Large-Scale Multimodal 3D Whole-body Human Motion Dataset". In: (2025). arXiv: 2501.05098 [cs.CV]. URL: https://arxiv.org/abs/2501.05098 (cit. on p. 5).