CoRe: A Hybrid Approach of Contact-aware Optimization and Learning for Humanoid Robot Motions

Taemoon Jeong[1†], Yoonbyung Chai[1†], Sol Choi[2,3], Jaewan Bak[2,4], Chanwoo Kim[1], Jihwan Yoon[1], Yisoo Lee[2], Kyungjae Lee[5], Joohyung Kim[6], Sungjoon Choi[1*]

***Abstract*— Recent advances in text-to-motion generation enable realistic human-like motions directly from natural language. However, translating these motions into physically executable motions for humanoid robots remains challenging due to significant embodiment differences and physical constraints. Existing methods primarily rely on reinforcement learning (RL) without addressing initial kinematic infeasibility. This often leads to unstable robot behaviors. We introduce Contact-aware motion Refinement (*CoRe*), a fully automated pipeline consisting of human motion generation from text, robot-specific retargeting, optimization-based motion refinement, and a subsequent RL phase enhanced by contact-aware rewards. This integrated approach mitigates common motion artifacts such as foot sliding, unnatural floating, and excessive joint accelerations prior to RL training, thereby improving overall motion stability and physical plausibility. We validate our pipeline across diverse humanoid platforms without task-specific tuning or dynamic-level optimization. Results demonstrate effective sim-to-real transferability in various scenarios, from simple upper-body gestures to complex whole-body locomotion tasks.**

I. Introduction

Designing expressive and executable motions for humanoid robots remains a labor-intensive process, typically relying on handcrafted trajectories, rule-based planners, or task-specific controllers. These approaches are difficult to scale or generalize across robots with different kinematics, environments, and tasks. As humanoid robots are increasingly deployed in interactive, human-centered settings, there is a growing need for automated and scalable methods that can translate high-level inputs into executable motions along with corresponding control strategies.

Recent advances in text-to-motion generation [1], [2] have demonstrated that diverse and semantically meaningful human motions can be synthesized from natural language using diffusion or transformer based generative models. While promising, these models are trained on human motion distributions and do not account for the kinematic or dynamic constraints of physical robots. As a result, directly applying text-generated human motions to real-world robots often leads to joint limit violations or damaging behaviors during execution [3].

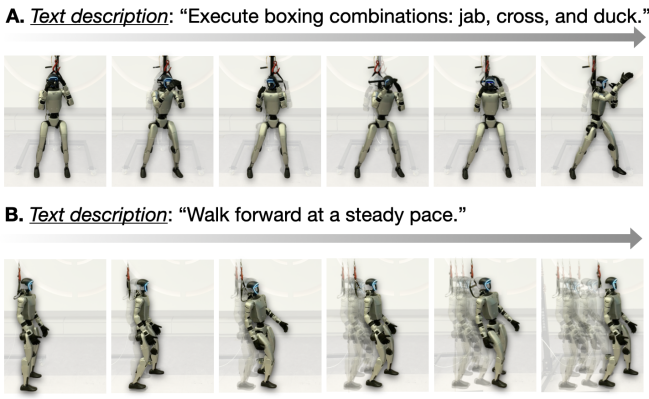


Fig. 1: **Real robot deployment.** G1 executing motions from text descriptions: (**A**) Boxing sequence, (**B**) Walking forward.

To address this challenge, we propose **Co**ntact-aware motion **Re**finement (***CoRe***), a fully automated pipeline that generates robot-executable motions from free-form text inputs. As illustrated in Fig. 1, our system allows real humanoid robots to perform both dynamic whole-body motion and locomotion from natural language commands. Our method begins by generating a human motion from natural language using a pre-trained generative model. The resulting motion is then retargeted to the target robot and refined kinematically to ensure stable and executable motion for the robot. This step ensures that the motion conforms to the robot's specific kinematic structure and exhibits stable foot-ground contact behavior. Finally, a physics-based reinforcement learning phase trains a low-level policy to imitate the refined motion in simulation, using both motion imitation and contact-aware rewards to ensure robustness and sim-to-real transferability.

A key feature of *CoRe* is the integration of kinematic refinement into the motion generation process. Unlike prior approaches [4]–[7] that rely solely on reinforcement learning to resolve physical constraints, *CoRe* performs a preemptive correction of motion artifacts—such as foot sliding, unnatural floating, or excessively high joint accelerations—that often arise when adapting human motions to robots. Furthermore, we incorporate the outcomes of kinematic refinement into the reward function of the imitation learning phase, allowing the policy to benefit from both geometric and dynamic feedback

[1]Taemoon Jeong, Yoonbyung Chai, Chanwoo Kim, Jihwan Yoon, and Sungjoon Choi are with the Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea (email: taemoon-jeong, yoonbyung-chai, yoonmungchi, sungjoon-choi@korea.ac.kr)

[2]Yisoo Lee, Sol Choi, and Jaewan Bak are with the Korea Institute of Science and Technology (KIST), Seoul, Republic of Korea (email: yisoo.lee, solchoi@kist.re.kr, jaewan.bak@kist.re.kr)

[3]Sol Choi is also with School of Mechanical Engineering, Yonsei University, Seoul, Republic of Korea (email: solchoi@yonsei.ac.kr)

[4]Jaewan Bak is also with School of Electrical Engineering, Korea University, Seoul, Republic of Korea (email: jaewan_bak@korea.ac.kr)

[5]Kyungjae Lee is with the Department of Statistics, Korea University, Seoul, Republic of Korea (email: kyungjae_lee@korea.ac.kr)

[6]Joohyung Kim is with the Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Champaign, IL, USA (email: joohyung@illinois.edu)

[†]These authors contributed equally to this work.

[*]Corresponding author

within a unified training framework.

Fig. 2 presents an overview of our proposed pipeline. Our contributions are summarized as follows:

1) We introduce *CoRe*, a contact-aware, kinematics-based motion refinement strategy that enhances the feasibility and deployability of retargeted motions without requiring dynamic-level optimization.
2) We propose a contact-guided reinforcement learning framework for motion imitation, enabling robust and stable policy execution in both simulation and on real hardware.
3) We validate our method across diverse robot embodiments and motion types—including whole-body locomotion and upper-body gestures—demonstrating its effectiveness and generality in real-world deployments.

II. Related Work

A. Robot Motion Retargeting

Robot motion retargeting adapts human motion to robots, managing morphological and kinematic discrepancies. Early foundational methods by Pollard et al. [8] provided initial frameworks for transferring motions between humans and robots. Choi and Kim [9] developed a structured pipeline for intuitive robotic motion generation, and subsequent data-driven methodologies such as the nonparametric approach by Choi et al. [10] and Self-Supervised Shared Latent Embedding (S3LE) [11] prioritized scalability and collision-free motion synthesis across various robot platforms. Jeong et al. [12] proposed a robust pipeline integrating a unified kinematic rig and trajectory refinement to handle diverse motion data.

B. Physics-based Motion Imitation

Peng et al. [13] introduced a reinforcement learning (RL) framework to train physically simulated agents in motion imitation tasks. Further advancements by Peng et al. [14] improved motion realism and stylistic diversity. Recent developments have emphasized continuous imitation and robustness, such as the perpetual motion controller by Luo et al. [15]. Hybrid methods combining simplified trajectory optimization with RL have shown further potential, as exemplified by Fuchioka et al. [16]. Ji et al. [6] proposed ExBody2, enhancing motion realism through a generalist-specialist policy structure but lacking seamless transitions between multiple specialized tasks. OmniH2O by He et al. [17] presented a whole-body humanoid teleoperation and autonomous control system that supports dexterous loco-manipulation but relies heavily on extensive sensor inputs and real-time teleoperation data. Additionally, He et al. [7] developed HOVER, a versatile neural controller consolidating diverse control modes, but its approach does not explicitly perform kinematic pre-refinement of generated motions, potentially limiting physical feasibility.

C. Text-based Robot Motion Generation

Recent works have explored the integration of large language models (LLMs) and vision-language models (VLMs) for robot motion generation from natural language descriptions. Kumar et al. [4] presented a method for iteratively refining robot control policies via natural language commands, enabling diverse humanoid robot behaviors without complex reward engineering; however, it lacks preemptive kinematic refinements to avoid joint violations and unstable behaviors. Similarly, Xu et al. [5] and LangWBC by Shao et al. [18] rely solely on RL methods without explicitly addressing initial kinematic infeasibility. RobotMDM by Serifi et al. [19] employed diffusion-based generative models fine-tuned with RL-based reward surrogates for plausible robot motions but required training separate models for each robot type, limiting scalability. Jiang et al. [20] introduced HARMON, utilizing human motion priors and VLMs to produce semantically meaningful motions validated through simulations and experiments, yet similarly does not integrate explicit kinematic adjustments for robot-specific constraints. In contrast, our method explicitly incorporates preemptive kinematic refinements to ensure motion executability and physical feasibility, addressing these limitations comprehensively.

III. Proposed Method

Our proposed method, **Co**ntact-aware motion **Re**finement (***CoRe***), introduces a unified pipeline that generates physically plausible and diverse humanoid motions directly from free-form text inputs. Unlike previous methods such as Words into Action [4] and LAGOON [5], which primarily focus on generative capabilities without addressing robot-specific constraints, *CoRe* explicitly integrates a kinematic-level optimization step to ensure physical feasibility of generated motions. Furthermore, compared to dynamic-level trajectory optimization approaches like OPT-mimic [16], our method significantly reduces computational complexity while preserving motion quality through targeted kinematic refinements. TABLE I highlights the distinctive features of *CoRe* relative to existing methods, emphasizing its unique combination of text-based synthesis, trajectory refinement, and computational efficiency.

TABLE I: **Qualitative comparison across methods.**
✓ : supported, ✗ : not supported, – : not applicable

Method	Text Input	Motion Diversity	Trajectory Optimization	Optimization Complexity
Proposed Method	✓	✓	✓	Low
OPT-mimic [16]	✗	✗	✓	High
Words into Action [4] LAGOON [5]	✓	✓	✗	–

A. Text-Conditioned Human Motion Generation

Our method starts by generating preliminary human motions from textual inputs using the Motion Diffusion Model (MDM) [1]. MDM synthesizes realistic and contextually relevant human motions based on semantic embeddings derived from textual descriptions. This step provides diverse initial motion sequences, which are subsequently refined to match robot-specific kinematic constraints.

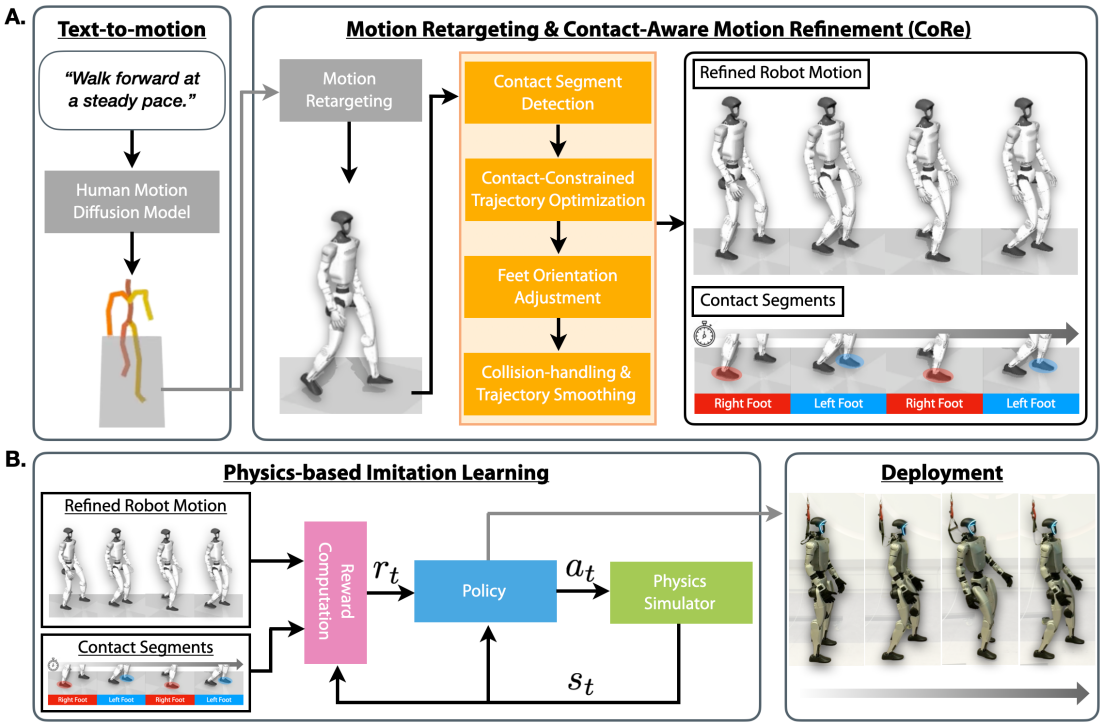


Fig. 2: **System overview.** (**A**) Our pipeline begins with text-to-motion generation from natural language, followed by robot-specific retargeting and Contact-aware motion Refinement (*CoRe*), which includes detecting stable contact segments, optimizing trajectories under contact constraints, adjusting feet orientations, and handling collisions. (**B**) The refined motion and extracted contact segments are utilized in physics-based imitation learning, where a reinforcement learning policy is trained with contact-aware rewards. This enables robust sim-to-real deployment, ensuring reliable and safe execution of robot motions corresponding to given text instructions.

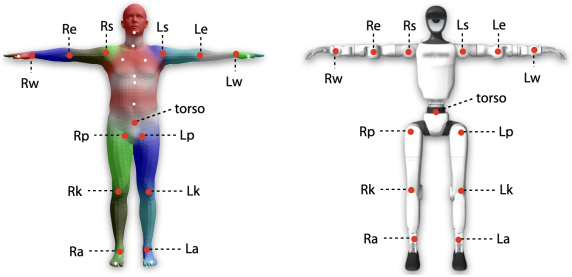


Fig. 3: **JOI information for SMPL and G1.** JOI includes key joints such as the head, shoulders, elbows, wrists, pelvis, knees, and ankles.

B. Robot Motion Retargeting

In this step, we transform generated human motions into executable robot motions using directional vectors. This involves identifying key Joints of Interest (JOI) [9], [12] specific to each robot's kinematic structure. Typically, JOI include major joints such as the head, shoulders, elbows, wrists, pelvis, knees, and ankles (Fig. 3). Due to kinematic differences between human and robot skeletons, direct joint mapping is not feasible. Instead, we compute scaled directional vectors from the human motion to define target positions for the robot's JOI, accommodating differences in limb lengths.

The target position of robot joint j is computed by:

$$\mathbf{p}_j^{\text{robot}} = \mathbf{p}_i^{\text{robot}} + l_{ij}^{\text{robot}} \cdot \mathbf{v}_{ij}^{\text{human}}$$

where l_{ij}^{robot} is the link length between robot joints i and j, and $\mathbf{v}_{ij}^{\text{human}}$ is the directional vector from joint i to j in the human skeleton.

These target positions serve as input to a numerical Inverse Kinematics (IK) solver. The IK formulation minimizes discrepancies between target joint transformations and those computed by Forward Kinematics (FK), subject to joint limit constraints:

$$\min_{\mathbf{q}_{1:N}^{\text{robot}}} \sum_{k=1}^{N} \left\| \hat{\mathbf{T}}_k^{\text{robot}} - \text{FK}(\xi^{\text{robot}}, \mathbf{q}_k^{\text{robot}}) \right\|^2$$
$$\text{s.t.} \quad \mathbf{q}_{\min} \leq \mathbf{q}_{1:N}^{\text{robot}} \leq \mathbf{q}_{\max}$$

where $\mathbf{q}_{1:N}^{\text{robot}}$ denotes joint angles over time, ξ^{robot} is the robot's kinematic model, and $\mathbf{q}_{\min}$, $\mathbf{q}_{\max}$ represent joint limits.

C. Contact-Aware Motion Refinement (CoRe)

Our proposed method refines the retargeted robot motions to ensure stable foot-ground interactions and overall physical feasibility. As illustrated in Fig. 4, *CoRe* systematically addresses critical issues such as unstable foot placements, unnatural joint orientations, and potential collisions through a structured kinematic optimization process. The refinement procedure includes robust detection of foot-ground contact segments, trajectory optimization constrained by these contacts, adjustment of foot orientations to ensure realistic interactions, and collision handling with trajectory smoothing.

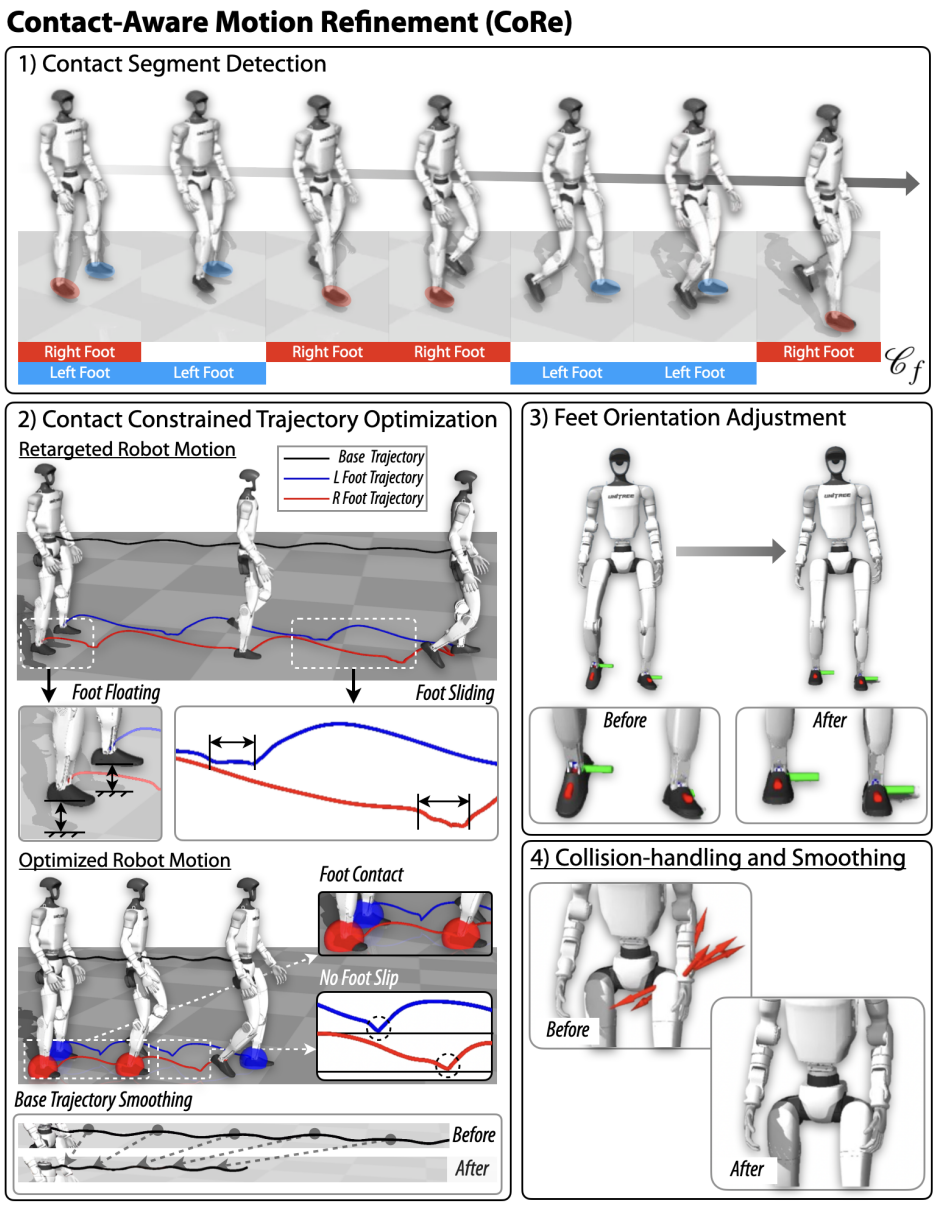


Fig. 4: **Contact-Aware Motion Refinement (*CoRe*)** 1) *Contact Segment Detection*: Identifying reliable foot-ground contacts ($\mathscr{C}_f$) by analyzing toe trajectories. 2) *Contact-Constrained Trajectory Optimization*: Refining trajectories to eliminate foot sliding and floating, ensuring stable ground interactions and smooth base motion. 3) *Feet Orientation Adjustment*: Optimizing foot yaw orientation to maintain natural and stable foot positioning during contacts. 4) *Collision-handling and Smoothing*: Resolving self-collisions through targeted position adjustments and smoothing trajectories to prevent abrupt changes.

This comprehensive optimization transforms initially retargeted trajectories into physically executable and dynamically consistent motions, significantly enhancing their suitability for deployment on real humanoid robots.

1) Contact Segment Detection: This step robustly identifies foot-ground contact segments, denoted as $\mathscr{C}_f$, by analyzing the vertical trajectories of the robot's toe joints. We first smooth the vertical (z-axis) trajectories to reduce measurement noise. Subsequently, vertical velocities are computed, and contact segments are detected when the absolute vertical velocity remains below a predefined threshold. Short transient contacts are excluded by enforcing a minimum duration for each segment. This detection is conducted independently for each foot, establishing reliable ground contact intervals essential for subsequent optimization stages and later utilized in defining the contact reward within the reinforcement learning framework.

2) Contact-Constrained Trajectory Optimization: This optimization step ensures the physical plausibility and stability of the robot's gait by refining the initial trajectories to comply with contact constraints, thus preventing slipping and unnatural movements. The optimization problem is defined as follows:

$$\min_{\hat{\mathbf{p}}_{1:N}} \quad \sum_{k=1}^{N} \Big(w_{\text{base}} \|\hat{\mathbf{p}}_k^{\text{base}} - \mathbf{p}_k^{\text{base}}\|^2 + w_{\text{feet}} \big(\|\hat{\mathbf{p}}_k^{\text{rf}} - \mathbf{p}_k^{\text{rf}}\|^2 + \|\hat{\mathbf{p}}_k^{\text{lf}} - \mathbf{p}_k^{\text{lf}}\|^2 \big) \Big)$$

$$\text{s.t.} \quad \hat{\mathbf{p}}_{z,k}^{f} \geq z_{\min}, \quad f \in \{\text{rf}, \text{lf}\} \tag{1a}$$

$$\Delta \hat{\mathbf{p}}_k^{f} = 0, \quad f \in \{\text{rf}, \text{lf}\}, \quad \forall k \in \mathscr{C}_f \tag{1b}$$

$$\|\hat{a}_k^{\text{base}}\| \leq a_{\max}^{\text{base}} \tag{1c}$$

$$\|\hat{a}_k^{f}\| \leq a_{\max}^{f}, \quad f \in \{\text{rf}, \text{lf}\} \tag{1d}$$

This formulation seeks the optimized positions $\hat{\mathbf{p}}$ by minimizing the weighted Euclidean distances between the initial and refined trajectories of the robot's base and feet. Here, w_{base} and w_{feet} represent weighting factors for prioritizing trajectory fidelity.

Each constraint serves a clear purpose: Constraint (1a) ensures that each foot remains above a minimal height $z_{\min}$ preventing ground penetration. Constraint (1b) enforces zero displacement for the feet during identified contact segments $\mathscr{C}_f$, preventing foot slippage. Constraints (1c) and (1d) limit accelerations of the base and feet, respectively, ensuring smooth trajectories without abrupt movements. By solving this optimization, we obtain physically consistent trajectories suitable for subsequent inverse kinematics calculations, resulting in natural and executable robot motions.

3) Feet Orientation Adjustment: This optimization step ensures stable and natural foot orientations during ground contact segments, $\mathscr{C}_f$. The method prevents unnecessary rotations and maintains consistent yaw orientations by formulating an optimization problem:

$$\min_{\hat{\mathbf{y}}_{1:N}} \quad \sum_{k=1}^{L} \Big(\|\hat{\mathbf{y}}_k^{\text{rf}} - \mathbf{y}_{\text{rf}}\|^2 + \|\hat{\mathbf{y}}_k^{\text{lf}} - \mathbf{y}_{\text{lf}}\|^2 \Big)$$

$$\text{s.t.} \quad \hat{\mathbf{y}}_k^{f} = \mathbf{y}_f, \quad f \in \{\text{rf}, \text{lf}\}, \quad \forall k \in \mathscr{C}_f \tag{2a}$$

$$\|\hat{\alpha}_z^{f}\| \leq \alpha_{z,\max}, \quad f \in \{\text{rf}, \text{lf}\} \tag{2b}$$

Here, $\hat{\mathbf{y}}_k^{f}$ denotes the optimized yaw trajectory of foot f, and $\mathbf{y}_f$ represents the initial yaw orientation. Constraint (2a) ensures the yaw orientation remains fixed during ground contacts, while constraint (2b) limits yaw angular acceleration, preventing abrupt orientation changes. Solving this optimization yields a stable yaw trajectory that significantly enhances the robot's stability and movement realism.

4) Collision Handling and Trajectory Smoothing: This refinement step addresses potential self-collisions by adjusting joint positions to ensure safe robot motion. Colliding body parts are shifted according to contact force vectors to resolve collisions effectively. Specifically, the target position of each colliding body is updated using:

$$\mathbf{p}_{\text{target}} = \mathbf{p}_{\text{contact}} + \lambda \cdot \mathbf{v}_{\text{contact}}$$

where is a small scalar controlling displacement magnitude along the direction of the contact force vector . After collision resolution, trajectories undergo smoothing by enforcing acceleration limits, thereby preventing abrupt changes and ensuring smoother, more stable robot motions.

D. Physics-Based Imitation Learning

To enable a humanoid to physically execute the refined motion, we employ a physics-based motion imitation framework based on reinforcement learning. The objective is to learn motor skills that can closely reproduce a given reference trajectory within a physically realistic simulation. In the physics simulation, a humanoid agent is trained to perform the motion by interacting with the environment and receiving higher rewards when its behavior closely matches the reference motions.

1) State and Action: At each timestep t, the policy receives an observation $o_t = [\mathbf{q}_t, \dot{\mathbf{q}}_t, \boldsymbol{a}_{t-1}, \mathbf{g}_t, \omega_t, \phi_t] \in \mathbb{R}^{95}$, which consists of information that can be reliably obtained from onboard sensors and estimators during real-world deployment. Specifically, the observation includes joint positions $\mathbf{q}_t \in \mathbb{R}^{29}$, joint velocities $\dot{\mathbf{q}}_t \in \mathbb{R}^{29}$, the previous action $\boldsymbol{a}_{t-1} \in \mathbb{R}^{29}$, the gravity vector projected into the base frame $\mathbf{g}_t \in \mathbb{R}^3$, the base angular velocity $\omega_t \in \mathbb{R}^3$, and a phase variable $\phi_t \in \mathbb{R}^2$ representing the progression of the reference motion. Based on the observation, the policy outputs the next action $\boldsymbol{a}_t \in \mathbb{R}^{29}$, which is the target position of the PD controller for the robot at the current timestep.

TABLE II: **Reward and penalty terms.** Reward terms including imitation objectives from reference motion and additional penalties for stable sim-to-real transfer

Index	Name	Weight
Rewards (from reference motion)		
0	joint position	1.95
1	joint velocity	0.30
2	root position	0.30
3	keypoint position	0.15
4	contact-aware	0.54
Penalty Terms (stability and regularization)		
5	motor torques	-2×10^{-8}
6	joint acceleration	-3×10^{-8}
7	action rate	-0.01
8	joint position limits	-5.00
9	feet slide	-0.20

2) Reward: The overall reward is composed of three components: imitation rewards, contact-aware rewards, and penalty terms. The overall reward is computed as a weighted sum of reward and penalty terms, with the corresponding weights provided in the TABLE II.

The imitation rewards are defined based on the discrepancy between the simulated motion and the reference motion, including differences in joint positions and velocities, as well as root position and orientation and keypoint position, following the standard formulation used in prior physics-based imitation learning methods.

Imitation Reward: First, joint position reward encourage the simulated agent to match the reference joint angles and global orientation of the root and the angles are represented in radians. The joint positions are compared in joint space,

and the root orientation is represented as a 3D rotation vector.

$$r_t^p = \exp\left[-2\left(\sum_j \left\|\hat{q}_t^j - q_t^j\right\|^2\right)\right]$$

where r_t^p denotes the reward for joint position tracking, and $\hat{q}_t$ and q_t represent the joint angles from the reference motion and the simulated trajectory, respectively, at time t.

Joint velocity reward penalizes discrepancies between the simulated and reference joint angular velocities.

$$r_t^v = \exp\left[-0.01\left(\sum_j \left\|\hat{\dot{q}}_t^j - \dot{q}_t^j\right\|^2\right)\right]$$

where r_t^v denotes the reward for joint velocity tracking, and $\hat{\dot{q}}_t$ and $\dot{q}_t$ represent the joint velocities from the reference motion and the simulated trajectory, respectively.

Root position reward measures the difference in the global position of the base (root) of the simulated agent and the base position of the reference.

$$r_t^{\text{base}} = \exp\left[-10\left(\left\|\hat{p}_t^{\text{base}} - p_t^{\text{base}}\right\|^2\right)\right]$$

where r_t^{base} denotes the reward for root position tracking, and $\hat{p}_t^{\text{base}}$ and p_t^{base} represent the root position from the reference motion and the simulated trajectory, respectively.

Keypoint position reward compares the 3D positions of selected keypoints (e.g., wrists and ankles) between the simulated and reference motions in world space, represented in the local coordinate frame of the pelvis base.

$$r_t^{\text{key}} = \exp\left[-40\left(\sum_{\mathbb{k}} \|\hat{p}_t^{\mathbb{k}} - p_t^{\mathbb{k}}\|^2\right)\right]$$

where $\mathbb{k} \in \{\text{rf}, \text{lf}, \text{rw}, \text{lw}\}$ denotes the set of keypoint bodies (right foot, left foot, right wrist, left wrist), and r_t^{key} denotes the reward for keypoint position tracking, and $\hat{p}_t^{\mathbb{k}}$ and $p_t^{\mathbb{k}}$ represent the keypoint position from the reference motion and the simulated trajectory, respectively.

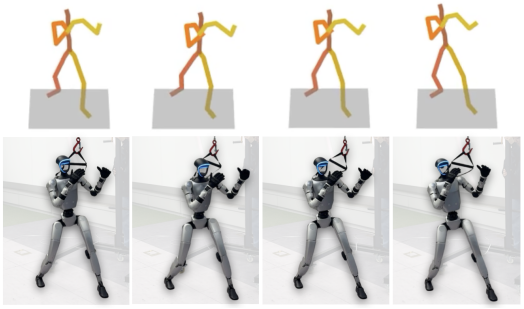
Contact Reward: To further improve motion fidelity, especially for dynamic behaviors like locomotion, we incorporate contact segments obtained during motion refinement into the reward function. These contact-aware terms help the agent achieve stable foot placement and ground contact pattern, improving both learning stability and physical plausibility.

$$\hat{c}_t^f = \begin{cases} 1 & \text{if } t \in \mathscr{C}_f \\ 0 & \text{otherwise} \end{cases}$$

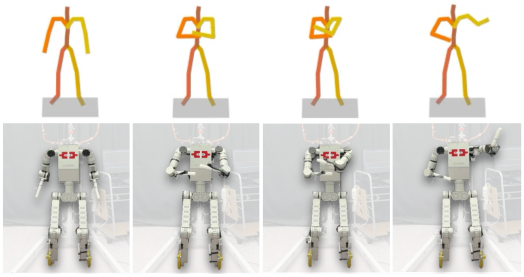
$$c_t^f = \mathbb{1}\left[\left\|F_t^f\right\| > \varepsilon\right]$$

$$r_t^{\text{contact}} = \sum_f \mathbb{1}\left[\hat{c}_t^f = c_t^f\right]$$

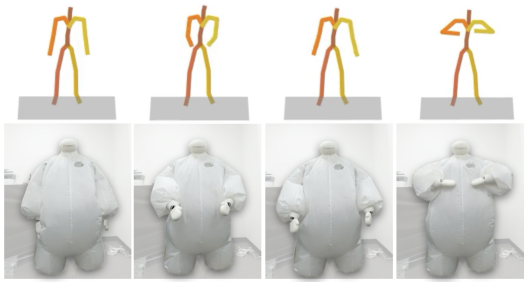
where $f \in \{\text{rf}, \text{lf}\}$ denotes the set of left and right feet, F_t^f represents the ground reaction force acting on foot f. and $\hat{c}_t$ and c_t indicate the reference contact segments and simulated contact signals, respectively. The threshold ε to determine contact is set to 1



(a) "Raise your arm to intercept an incoming strike, using a blocking motion."



(b) "Cross your arms over your chest confidently starting in neutral pose."



(c) "Open arms for friendly hug."

Fig. 5: **Sanpshots of real-robot deployment**

Penalty Term: To promote stable and realistic motion, we introduce several penalty terms to the reward function. These include L2 norm on motor torque and joint acceleration to reduce energy usage and motion jitter, and an action rate penalty to encourage smoother transitions between actions. A joint limit penalty is applied when joint positions exceed 90% of their valid range, preventing unnatural configurations. Additionally, a foot sliding penalty is imposed when feet are in contact with the ground but exhibit sliding motion.

IV. Experiment

A. Experimental Setup

1) Data Preparation: We generated a total of 90 motions based on text descriptions, with 30 motions for each of the following three categories:

- **Simple Stationary Motions**: Basic motions without base movement or self-collisions (e.g., "Wave your hand to say hello").
- **Walking Motions**: Motions involving ground contact and base movement (e.g., "Walk forward at a steady pace").
- **Complex Motions with Self-Collision**: Dynamic motions with high potential for self-collisions (e.g., "Cross

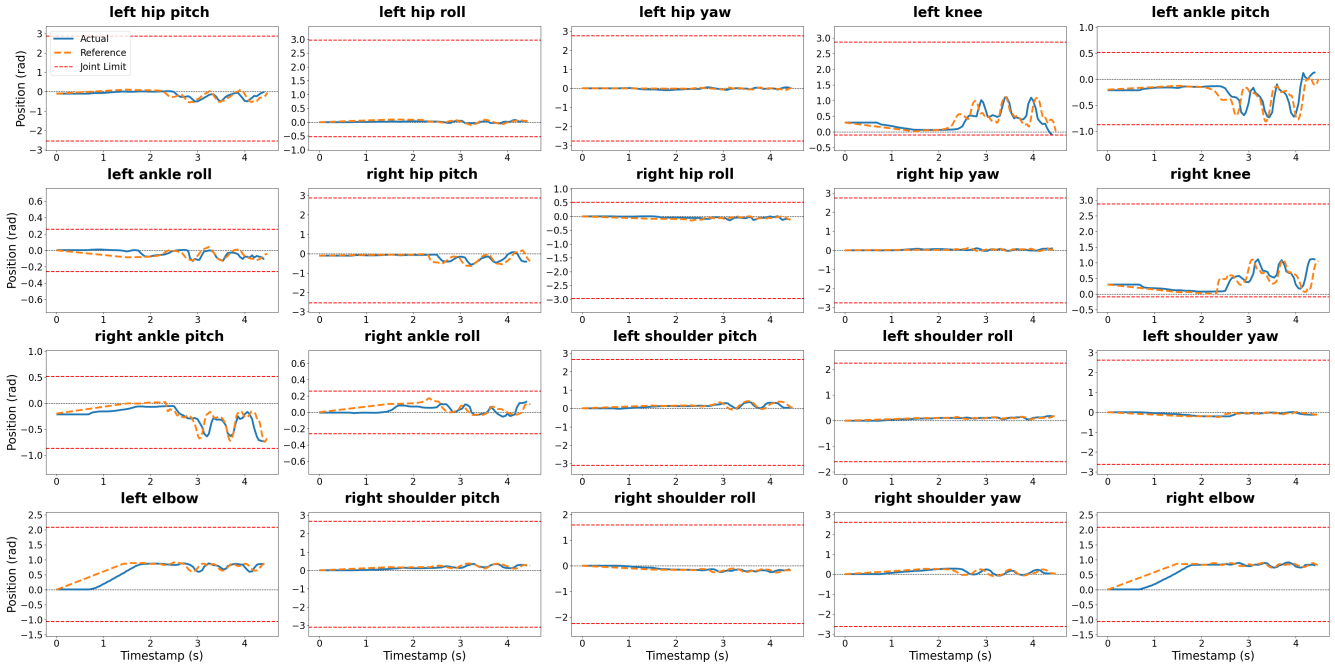


Fig. 6: **Tracking performance of G1 during real-world locomotion.** Actual joint positions (solid lines) closely follow the reference trajectories (dashed lines), demonstrating accurate motion reproduction. The actual motion shows a slight phase shift relative to the reference, but the whole-body trajectory accurately reproduces the reference motion throughout the walking sequence.

your arms over your chest confidently starting in neutral pose").

2) Evaluation Metrics: We assess our method's performance using a success rate metric, comparing the final learned motion with the reference motion. A motion is considered successful if it meets two key criteria. First, the robot must maintain stability throughout the motion without falling. Second, the robot must execute at least 90% of the motion duration without violating the stability criterion. The success rate is calculated as the percentage of successful motions within each category.

We chose these two criteria to highlight our approach's key advantages. Direct application of human motions to robots often fails due to kinematic differences. Our stability criterion shows our method's effectiveness in adapting motions for robot balance, while the completion criterion demonstrates fidelity to the original intent. These metrics together illustrate our method's capability to transform potentially unsuitable source motions into executable and stable robot actions, addressing a core challenge in text-to-robot motion generation.

B. Simulation Results

We compare our method with two baselines: Words into Action [4] and LAGOON [5]. TABLE III presents the success rate comparison across the three motion categories. Each generated motion sequence has a duration of 6 seconds and incorporates both locomotion and gestural components.

The results presented in TABLE III showcase our method's performance, particularly its strengths in complex motion categories. For Simple Stationary Motions, all methods achieve a perfect success rate of 100%, indicating that these basic motions are well-handled by existing approaches as well as our method.

TABLE III: **Success Rate Comparison (%) across Motion Categories**

Method	Simple Stationary Motions	Walking Motions	Complex Motions with Self-Collision
Proposed Method	**100.0**	**73.3**	**66.7**
Words into Action [4]	100.0	23.3	6.7
LAGOON [5]	100.0	26.7	20.0

The advantages of our approach become evident in more challenging scenarios. In the Walking Motions category, our method outperforms the baselines with a success rate of 73.3%, compared to Words into Action (23.3%) and LAGOON (26.7%). This improvement can be attributed to our motion refinement step, which effectively handles the physical constraints of ground contact and balance during locomotion.

Our approach also shows notable gains in the Complex Motions with Self-Collision category. In this challenging scenario, our method achieves a success rate of 66.7%, considerably higher than both Words into Action (6.7%) and LAGOON (20.0%). Our refinement process effectively manages self-collisions, allowing the robot to maintain balance and complete the motion. In contrast, due to the physical implausibility of reference motions, baseline methodologies frequently result in robot instability or failure to execute

intended kinematic sequences.

C. Real Robot Deployment

To evaluate the practicality of our proposed method, we conducted experiments with three humanoid robots: one upper-body robot and two full-body robots. The experiments demonstrate the successful deployment of text-generated motions on humanoid robots with diverse embodiments. Fig.5 presents snapshots of various robots executing motions synthesized from natural language commands. First, the full-body humanoid robot demonstrates stable lower-body support during upper-body motions, such as "Raise your arm to intercept an incoming strike, using a blocking motion" (Fig.5(a)). Both feet remain grounded, enabling the robot to perform expressive gestures while maintaining balance. Next, the wheeled humanoid robot executed a whole-body motion generated from the instruction "Cross your arms over your chest confidently starting in neutral pose," demonstrating the system's ability to produce balanced and coordinated whole-body behaviors from free-form text input (Fig.5(b)). Finally, the upper-body humanoid robot platform successfully performed upper-body gestures in response to instructions such as "Open arms for friendly hug" (Fig.5(c)).

Fig. 6 illustrates the motion tracking performance of G1, comparing the actual and desired joint positions. The ankle joint trajectories closely follow the orientation adjustments applied during the refinement process. Furthermore, contact-guided reward plays a critical role in stabilizing the gait pattern by ensuring that each foot makes contact with the ground and lifts off at the appropriate phase, preventing stumbling and maintaining stable foot-ground interaction. The results underscore the robustness and stability of our approach in real-world deployments.

V. Conclusion

In this paper, we introduced *CoRe*, a hybrid method integrating contact-constrained kinematic optimization and reinforcement learning-based fine-tuning, effectively enabling robots to perform physically plausible motions generated from natural language commands. Experimental results across diverse motion categories demonstrated significant improvements over baseline approaches, notably in complex scenarios involving locomotion and self-collisions. The practicality and versatility of our framework were validated through successful deployments on multiple humanoid robots, including full-body humanoids, upper-body humanoids, and wheel-based platforms. Thus, this work substantially contributes to bridging the gap between natural language and executable robot motions, with potential applications in assistive robotics, entertainment, and collaborative tasks.

Despite promising results, our method shows limitations in highly dynamic situations, such as rapid directional changes during locomotion or fast-paced actions like sprinting. Future work will therefore focus on enhancing robustness and real-time stability, particularly addressing challenges in dynamic and agile motion execution.

References

[1] 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.

[2] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8255–8263, 2023.

[3] Yashuai Yan, Esteve Valls Mascaro, Tobias Egle, and Dongheui Lee. I-ctrl: Imitation to control humanoid robots through constrained reinforcement learning, 2025.

[4] K Niranjan Kumar, Irfan Essa, and Sehoon Ha. Words into action: Learning diverse humanoid robot behaviors using language guided iterative motion refinement. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[5] 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)*, pages 9743–9750. IEEE, 2024.

[6] 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.

[7] 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. *arXiv preprint arXiv:2410.21229*, 2024.

[8] N.S. Pollard, J.K. Hodgins, M.J. Riley, and C.G. Atkeson. Adapting human motion for the control of a humanoid robot. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 2, pages 1390–1397 vol.2, 2002.

[9] Sungjoon Choi and Joohyung Kim. Towards a natural motion generator: a pipeline to control a humanoid based on motion data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4373–4380, 2019.

[10] Sungjoon Choi, Matthew KXJ Pan, and Joohyung Kim. Nonparametric motion retargeting for humanoid robots on shared latent space. In *Robotics: Science and Systems*, 2020.

[11] Sungjoon Choi, Min Jae Song, Hyemin Ahn, and Joohyung Kim. Self-supervised motion retargeting with safety guarantee. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8097–8103. IEEE, 2021.

[12] Taemoon Jeong, Taehyun Byun, Jihoon Kim, Keunjoon Choi, Jaesung Oh, Sungpyo Lee, Omar Darwish, Joohyung Kim, and Sungjoon Choi. Robust robot motion retargeting: Rig unification and application to diverse robots. 2024.

[13] 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.

[14] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4), July 2021.

[15] 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*, pages 10895–10904, 2023.

[16] Yuni Fuchioka, Zhaoming Xie, and Michiel Van de Panne. Opt-mimic: Imitation of optimized trajectories for dynamic quadruped behaviors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5092–5098, 2023.

[17] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[18] 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.

[19] 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*, pages 1–9, 2024.

[20] Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. *arXiv preprint arXiv:2410.12773*, 2024.