# **Whole-Body UMI**: Transferring UMI Manipulation Skills to Humanoid Whole-Body Manipulation via Real-Time Motion Generation

Anonymous Authors

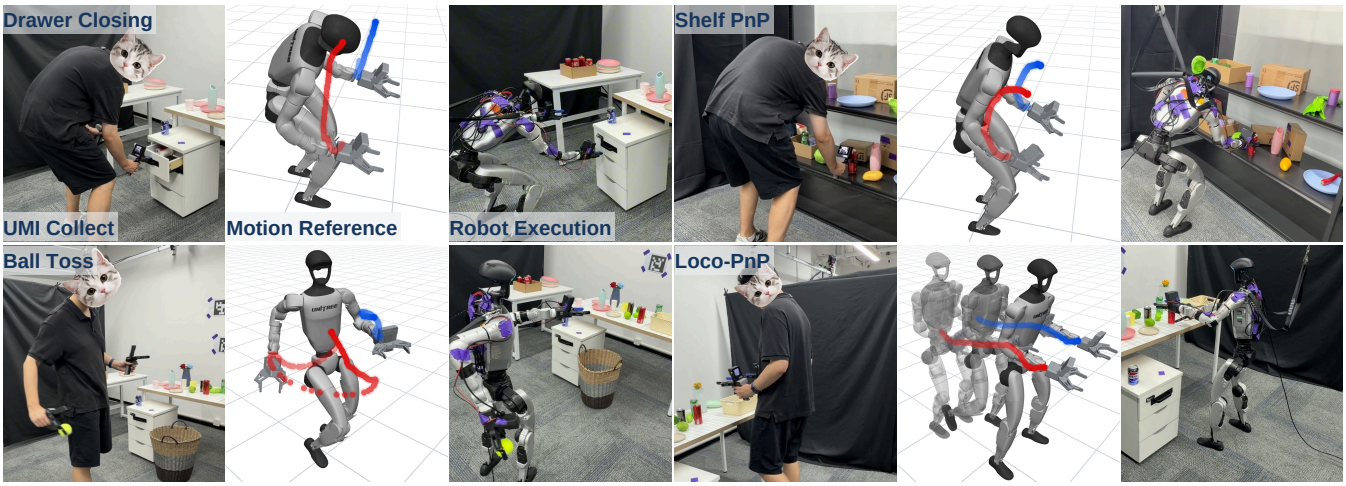


Fig. 1. **Whole-Body UMI: Universal Manipulation Interface with a Whole-Body Motion Prior**. We show the UMI manipulation skill transfer process for four real-robot tasks: drawer closing, shelf pick-and-place (PnP), ball toss, and locomotion pick-and-place (Loco-PnP). For each task, the panels from left to right show UMI end-effector data collection, generated whole-body references, and real-robot execution. The hierarchy enables a real humanoid to perform whole-body manipulation, including dynamic and locomotion-based tasks, using skills learned from native UMI demonstrations.

***Abstract*— Collecting whole-body demonstrations for humanoid manipulation mostly relies on teleoperation, which is costly and hard to scale up. The Universal Manipulation Interface (UMI) provides a scalable data collection paradigm, but end-effector trajectories alone underdetermine humanoid whole-body coordination, which is insufficient for whole-body demonstration collection. Therefore, we introduce Whole-Body UMI (WB-UMI), a task-agnostic, real-time and end-effector conditioned motion generator that decouples whole-body coordination learning from task semantics learning through a shared end-effector interface. A diffusion policy learns from native UMI demonstrations, while WB-UMI learns independently from retargeted motion capture, requiring no body trackers or paired image–whole-body demonstrations during task-specific data collection. In real deployment, an asynchronous hierarchy integrates the diffusion policy, motion generator, and a whole-body controller with latency compensation and measured-state feedback. Real-robot experiments on G1 support real-time closed-loop transfer across four tasks, achieving 90% success in drawer closing, 80% in shelf pick-and-place, 30% in ball toss, and 40% in Loco-PnP, which shows the effectiveness of this hierarchy in transferring native UMI skills to humanoid whole-body manipulation.**

## I. Introduction

Humanoid whole-body manipulation is valuable for many tasks in human environments. Achieving precise and natural-like humanoid whole-body manipulation requires whole-body coordination data, consisting of task-centric inputs (e.g., wrist-camera images to end-effector trajectories) and complex action spaces (e.g., movements of the arms, torso, and legs). One of the most effective ways is to learn from human demonstrations. However, most existing human data collection paradigms use teleoperation to obtain whole-body coordination through task-specific whole-body demonstrations, which is laborious, costly, and hard to scale up. The Universal Manipulation Interface (UMI) enables scalable, robot-free collection of task-centric end-effector (EE) demonstrations [1], offering a promising way to reduce this data collection burden.

However, EE trajectories specify task intent but fundamentally underconstrain humanoid whole-body coordination: the same EE trajectories can be realized through different body postures and foot placements. How to generate plausible and effective whole-body coordination from UMI data therefore still remains an open challenge. Existing solutions add body trackers to UMI and then use inverse kinematics (IK) with hand-designed constraints to obtain coordinated whole-body motion [2], [3]. This process provides direct coordination supervision, but introduces two limitations: *(1) Additional instrumentation:* Users must wear and calibrate multiple trackers in addition to the handheld UMI device, compromising the convenience of UMI. *(2) Limited scalability:* Although UMI with trackers improves collection efficiency over teleoperation, it still requires task-specific whole-body demonstrations. This dependence remains a fundamental bottleneck to scaling data collection for rich and diverse whole-body coordination.

*Can we leverage UMI's robot-free collection paradigm to avoid costly whole-body demonstrations and enable scalable humanoid whole-body manipulation at the same time?* Our key idea is to decouple task semantics learning from whole-

body coordination learning through a shared EE trajectory interface. Therefore, we introduce **Whole-Body UMI (WB-UMI)**, a task-agnostic real-time and EE-conditioned motion generator trained on mocap independently. In our hierarchy (Fig. 2(a)), a diffusion policy (DP) trained on native UMI demonstrations predicts EE trajectories. **WB-UMI** combines these trajectories with fused whole-body history to generate explicit whole-body references for a generalist whole-body controller (WBC). This proposed hierarchy connects task semantics learning to whole-body execution without additional task-specific paired whole-body motion collection and is validated across four tasks, achieving real-time, closed-loop execution (Fig. 1). In summary, our contributions are as follows:

*(1) A new framework that decouples task semantics learning from whole-body coordination learning:* We decouple scalable UMI data collection from costly humanoid demonstrations by learning whole-body coordination independently from existing mocap and connecting it to task-specific skill learning through a shared end-effector (EE) interface. This preserves UMI's convenience while supporting scalable learning of rich humanoid whole-body coordination.

*(2) Novel design of WB-UMI, a task-agnostic, real-time and EE-conditioned whole-body motion generator for humanoids:* **WB-UMI** enables whole-body motion generation from EE-only targets, without additional body-keypoint constraints. It learns a rich, task-agnostic whole-body motion prior from existing mocap that generalizes to unseen action categories (Table V). It further combines low-latency streaming generation with measured-state feedback from the robot for real-time deployment.

*(3) A novel three-layer system for asynchronous closed-loop deployment:* We integrate the task-specific DP, **WB-UMI**, and WBC for real-time execution from images to whole-body actions. The system compensates for generation latency, coordinates replanning through controller feedback, and preserves manipulation targets during robot motion (Fig. 4). In simulation, measured-state feedback improves EE orientation and whole-body reference tracking (Table VI). On the real robot, the integrated system achieves success rates of 90%, 80%, 30%, and 40% on drawer closing, shelf PnP, ball toss, and Loco-PnP, respectively (Fig. 5).

## II. Related Work

In this section, we review UMI data collection, conditioned motion generation, and Woble-body Control (WBC).

### A. UMI and the Humanoid Whole-Body Gap

Demonstration collection for humanoid manipulation differs in how it acquires task semantics and whole-body coordination. Teleoperation records paired image–whole-body coordination under human guidance [4]–[6], yielding robot-native demonstrations. However, collecting such data is laborious, costly, and hard to scale up to diverse environments. UMI [1], [7], instead, collects wrist-mounted images and EE trajectories directly from human demonstrations, offering a promising approach to addressing scalability challenges.

TABLE I: Motion generator capabilities. EE-only: separate world-space position/rotation constraints without extra future body targets. Real-time: reported interactive generation. Stream.: continuous generation of motion segments. FB: measured-state feedback. *Spatial control through inference-time optimization.

| | Online | | EE-only | | Robot | | Context (s) | |
|---|---|---|---|---|---|---|---|---|
| Method | Real-time | Stream. | Pos. | Rot. | FB | Real robot | Hist. | Future |
| MaskControl (full) [15] | × | × | ✓* | × | × | × | – | 10.00 |
| Kimodo [16] | × | × | × | ✓ | × | × | – | 10.00 |
| DartControl [17] | ✓ | ✓ | ✓* | × | × | × | 0.07 | 0.27 |
| ARDY [18] | ✓ | ✓ | × | ✓ | × | × | 8.00 | 10.00 |
| **WB-UMI (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.27 | 2.13 |

This robot-free collection paradigm has been extended to quadruped manipulator controllers [8], as well as to bimanual mobile manipulator [9], which proves the effectiveness of the UMI interface. However, for high-DoF floating-base humanoid robots [10], EE trajectories alone leave whole-body motion underconstrained. Recent robot-free interfaces address this gap by recording pelvis and foot motion alongside EE trajectories, then retargeting these sparse keypoints through inverse kinematics (IK) [2], [3]. These interfaces retain robot-free collection while supplying body targets for retargeting and control. However, their coordination data must still be collected alongside task demonstrations. In contrast, WB-UMI learns a reusable whole-body motion prior from existing mocap, thereby decoupling whole-body coordination learning from task-specific UMI demonstrations while preserving UMI's native EE-only collection interface.

### B. Conditioned Motion Generation

Generative models learn reusable whole-body motion priors from mocap, supporting diverse synthesis and adaptation to new conditions [11], [12]. Spatial control builds on these priors to make generated motion follow specified targets. Guidance during diffusion sampling enforces global trajectories or sparse joint-position constraints [13], [14]. Controllability has also been extended to masked motion models with sparse and dense joint-position constraints [15]. Directly conditioned models incorporate position and orientation targets into generation, supporting flexible keyframes and trajectories for motion authoring [16].

Temporal continuity poses a complementary challenge for online use. Autoregressive generators synthesize successive short motion segments from history to accommodate changing commands [17]. Larger-context models additionally condition on future goals beyond the current prediction window [18]. Recent humanoid systems combine language-conditioned streaming generation with low-level tracking for real-robot execution [19]. For UMI transfer, the generator must follow predicted EE position and orientation trajectories while adapting to execution errors. **WB-UMI** combines native EE-only conditioning, sparse look-ahead context, and measured-state feedback to generate explicit whole-body references online, without additional future body-keypoint targets (Table I).

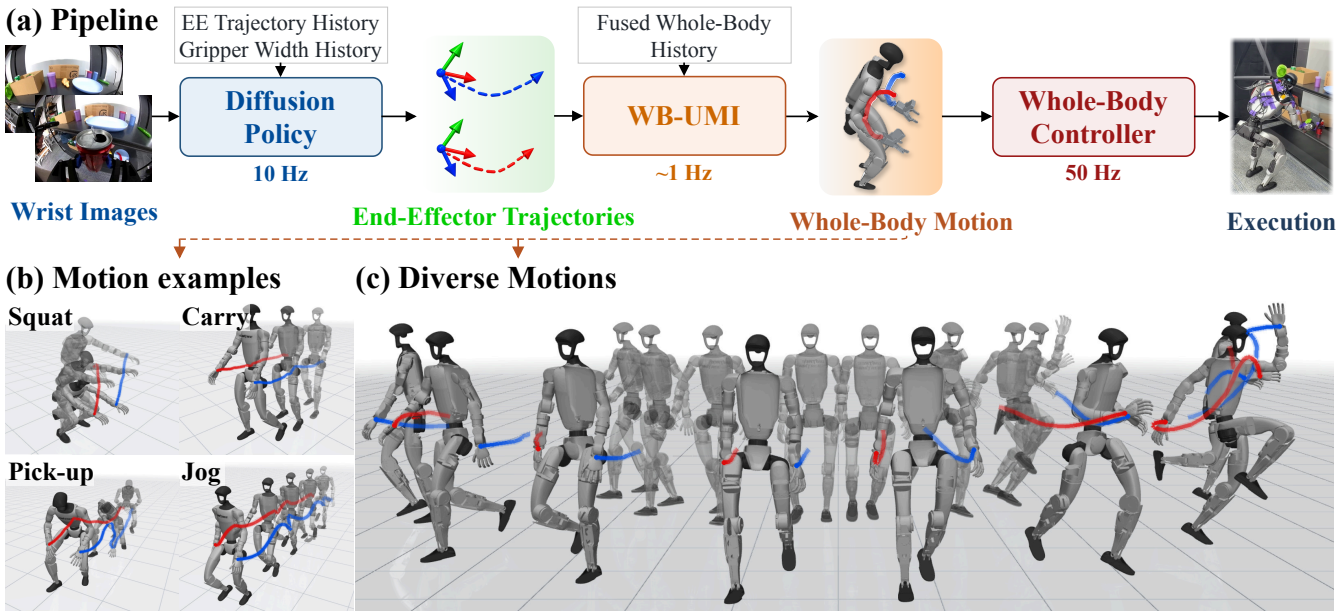


Fig. 2. Closed-loop deployment and generated motions. **(a)** A 10-Hz DP supplies EE trajectories to **WB-UMI** (∼1 Hz), whose references are tracked by a 50-Hz WBC. **(b)** Squat, carry, pick-up, and jog motions. **(c)** Motions generated from EE trajectories at different speeds. Red/blue curves mark EE trajectories. The EE interface supports diverse whole-body coordination within a closed-loop visual control hierarchy.

### C. Whole-Body Control

Classical whole-body control (WBC) often relies on hierarchical quadratic programming (QP) to coordinate multiple tasks under kinematic and dynamic constraints [20], [21]. Reinforcement learning (RL)-based motion trackers have been developed for simulated characters and humanoid robots [22]–[24]. Scaling motion data and policy capacity further enables generalist WBC that tracks unseen motions and supports multiple command interfaces with a single policy [25]. Alongside generalist tracking, behavior foundation models broaden task specification through shared latent representations of motions, goals, and rewards [26] and unified motion-command interfaces [27]. In contrast, **WB-UMI** converts EE-only UMI commands into explicit whole-body references for a pretrained generalist WBC.

## III. Method

In this section, we introduce our method in detail, including the problem formulation, **WB-UMI**, and description of its integration into a closed-loop deployment system.

### A. Problem Formulation and System Overview

We assume access to native UMI task demonstrations, an independent dataset of retargeted mocap, and a whole-body controller, but no task-specific whole-body demonstrations. Given observation history $\mathbf{o}_t$, comprising wrist RGB images and proprioceptive measurements, we seek a closed-loop system that generates coordinated whole-body actions in real time, enabling the humanoid to achieve the task-specific whole-body manipulation with only end-effector-level task demonstrations, where no future other body-keypoint targets are provided. Figure 2(a) illustrates our three-layer system. A DP trained on native UMI demonstrations maps $\mathbf{o}_t$ to EE trajectories and gripper widths. **WB-UMI**, trained independently on mocap, combines these EE trajectories with fused whole-body history to generate explicit whole-body kinematic references. The pretrained SONIC [25], which serves as our WBC here, tracks these references to execute the action. Controller feedback updates the observation and whole-body histories for subsequent predictions.

### B. *WB-UMI*: Real-Time EE-Conditioned Generation

Given whole-body history and EE future trajectories, **WB-UMI** infers coordinated whole-body motion without additional future body-keypoint targets. We autoregressively generate long-horizon motion as a sequence of short, fixed-length primitives [19], enabling continuous real-time updates in response to evolving EE targets. Each primitive contains $F$ future frames conditioned on $H$ history frames:

$$\mathbf{S}^{\mathrm{body},+} = G_\theta\left(\mathbf{S}^{\mathrm{body},-}, \mathbf{S}^{\mathrm{ee}}, \mathbf{L}^{\mathrm{ee}}\right), \tag{1}$$

where $\mathbf{S}^{\mathrm{ee}}$ supplies EE conditions aligned with $\mathbf{S}^{\mathrm{body}} = [\mathbf{S}^{\mathrm{body},-}, \mathbf{S}^{\mathrm{body},+}]$, and $\mathbf{L}^{\mathrm{ee}}$ provides sparse look-ahead context. We write $\mathbf{c} = (\mathbf{S}^{\mathrm{ee}}, \mathbf{L}^{\mathrm{ee}})$. Two primitives form a $2F$-frame motion segment. Table II defines the notation, and Fig. 3 summarizes the model.

**Data Construction.** For training, we construct paired whole-body states $\mathbf{s}_i^{\mathrm{body}}$ and EE conditions $\mathbf{s}_i^{\mathrm{ee}}$ entirely from retargeted G1 mocap [28], extracting EE trajectories through forward kinematics (FK). Table III summarizes both representations.

The whole-body state $\mathbf{s}_i^{\mathrm{body}}$ builds on a local incremental robot representation [19], making frame-to-frame changes

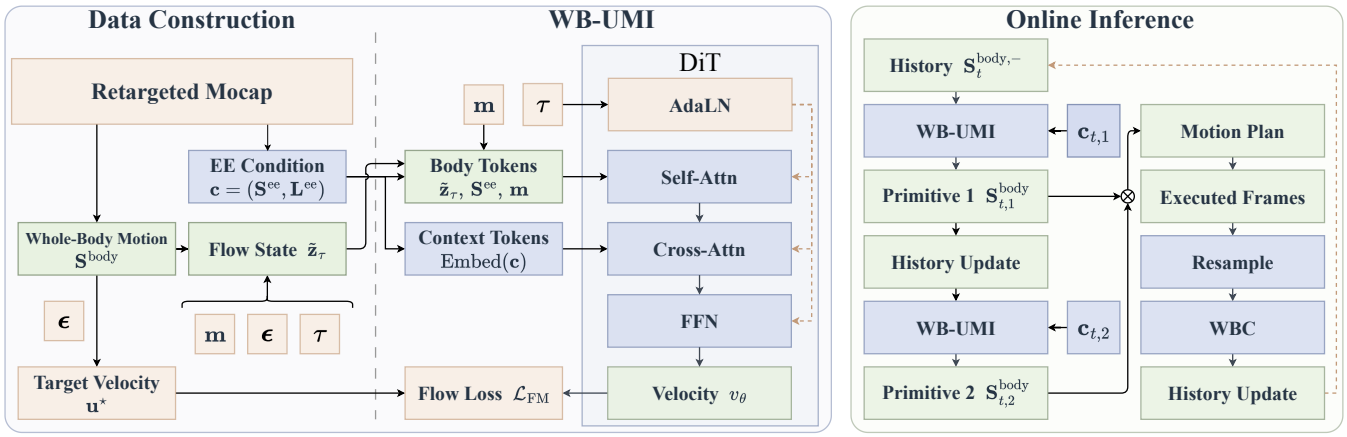


Fig. 3. **WB-UMI learns an EE-conditioned whole-body motion prior from retargeted mocap and deploys it through closed-loop autoregressive generation.** Left: Retargeted mocap provides paired whole-body motions and EE conditions for training. Center: The **WB-UMI** architecture is trained with masked flow matching to predict an EE-conditioned velocity field. Right: Detailed inference process for a single motion segment with execution history feedback.

TABLE II: Core notation for **WB-UMI**.

| Notation | Description | Value |
|---|---|---|
| $-$ / $+$ | History / future segment | – |
| $i, t, k$ | Frame (30 Hz), replanning step, and primitive index ($k \in \{1, 2\}$) | – |
| $H$ / $F$ | History / prediction length (frames) | 8 / 32 |
| $K$ / $d$ | Look-ahead length / sampling stride (frames) | 32 / 4 |
| $\text{Sample}_d$ | Stride-$d$ sampling from the first frame; 8 EE samples per look-ahead window | – |
| $\mathbf{c}_{t,k}$ | EE condition for primitive $k$ at step $t$ | – |
| $\mathbf{S}_{t,k}^{\text{body}}$ | Generated $F$-frame primitive $k$ at step $t$ | – |

explicit while retaining absolute root height. The EE condition $\mathbf{s}_i^{\text{ee}}$ combines relative poses, pose increments, and initial height and orientation anchors. Both representations are anchored at the first history frame of each segment. Root planar motion and EE displacements are measured relative to their respective initial positions and rotated by the inverse initial root yaw. EE rotations are relative to their initial orientations, and orientation anchors are expressed in the initial root heading frame. This local formulation reduces dependence on global placement.

TABLE III: Representations of the whole-body state and EE condition.

| Component | Representation | Size |
|---|---|---|
| *Whole-body state* $\mathbf{s}_i^{\text{body}}$ | | **Sum: 73** |
| Root orientation | Roll/pitch sine–cosine; yaw increment | 5 |
| Root position | Translation increment, height, planar pose | 8 |
| Joint motion | Joint positions and increments | 58 |
| Foot contact | L/R: ankle $h < 0.08$ m, $v < 0.1$ m/s | 2 |
| *EE condition* $\mathbf{s}_i^{\text{ee}}$ | | **Sum: 50** |
| Relative pose | Segment-relative position and 6D orientation | 18 |
| Height anchor | Initial heights | 2 |
| Orientation anchor | Initial heading-frame 6D orientations | 12 |
| Pose increment | Position and orientation increments | 18 |

At inference, the DP provides EE trajectories extending beyond the current primitive. The segment aligned with the generated frames serves as tracking targets, while less reliable farther predictions are sparsely sampled as look-ahead context to convey motion intent. The resulting EE condition is:

$$\begin{aligned} \mathbf{S}^{\text{ee}} &= (\mathbf{s}_i^{\text{ee}})_{i=1}^{H+F}, \\ \mathbf{L}^{\text{ee}} &= \text{Sample}_d\big((\mathbf{s}_i^{\text{ee}})_{i=H+F+1}^{H+F+K}\big). \end{aligned} \tag{2}$$

**Training.** We train **WB-UMI** on pairs of consecutive primitives using masked conditional flow matching. During training, we gradually replace the second primitive's ground-truth history with generated motion from the first. For each primitive, we construct the input at flow time $\tau \in [0, 1]$ using Gaussian noise $\boldsymbol{\epsilon}$:

$$\tilde{\mathbf{z}}_\tau = \big[\mathbf{S}^{\text{body},-},\ (1-\tau)\boldsymbol{\epsilon} + \tau \mathbf{S}^{\text{body},+}\big], \tag{3}$$

where the history remains fixed, while the future segment transitions linearly from noise at $\tau = 0$ to ground-truth motion at $\tau = 1$. The binary mask $\mathbf{m}$ marks history frames with 1 and future frames with 0. The conditional DiT $v_\theta$ predicts flow velocity using two EE conditioning paths (Fig. 3). It forms each body token from a frame of $\tilde{\mathbf{z}}_\tau$ and the time-aligned EE features in $\mathbf{S}^{\text{ee}}$. Self-attention models temporal dependencies among these tokens, while cross-attention provides access to the full EE context $\mathbf{c}$. For auxiliary supervision, we use the predicted velocity to estimate the clean future:

$$\widehat{\mathbf{S}}^{\text{body},+} = \tilde{\mathbf{z}}_\tau^+ + (1-\tau) v_\theta^+(\tilde{\mathbf{z}}_\tau, \tau, \mathbf{c}). \tag{4}$$

The flow-matching loss supervises each primitive's future velocity against the target flow velocity $\mathbf{u}^\star = \mathbf{S}^{\text{body},+} - \boldsymbol{\epsilon}$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\big\| v_\theta^+(\tilde{\mathbf{z}}_\tau, \tau, \mathbf{c}) - \mathbf{u}^\star \big\|_2^2, \tag{5}$$

where the expectation is over training windows, flow times, and noise. For all auxiliary losses below, $\mathbf{S}^{\text{body},+}$ and $\widehat{\mathbf{S}}^{\text{body},+}$ denote the concatenated ground-truth futures and clean estimates of both primitives. EE targets are frame-aligned, and geometric comparisons use a common coordinate frame. The state loss combines reconstruction and root consistency:

$$\mathcal{L}_{\text{state}} = \|\widehat{\mathbf{S}}^{\text{body},+} - \mathbf{S}^{\text{body},+}\|_2^2 + \ell_{\text{root}}, \tag{6}$$

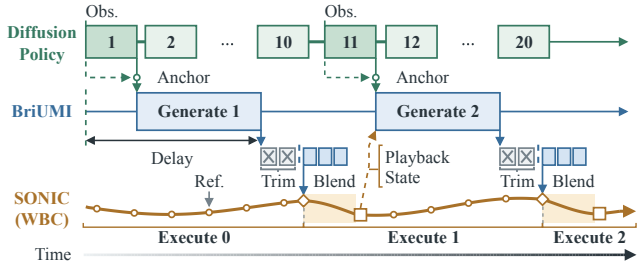


Fig. 4. Asynchronous deployment. **WB-UMI** generates new motion while the WBC tracks the active reference. Incoming references are trimmed and blended based on playback progress.

where $\ell_{\text{root}}$ encourages accurate root placement and internally consistent root motion. The EE loss compares FK-derived relative poses with the frame-aligned targets:

$$\mathcal{L}_{\text{EE}} = \|\widehat{\mathbf{S}}^{\text{ee}}_{\text{pose}} - \mathbf{S}^{\text{ee}}_{\text{pose}}\|_2^2 + \ell_{\text{rot}}(\widehat{R}, R_c), \tag{7}$$

where $\ell_{\text{rot}}$ penalizes orientation errors between predictions $\widehat{R}$ and targets $R_c$. We supervise body poses, velocities, and accelerations with

$$\mathcal{L}_{\text{coord}} = \sum_{i=0}^{2} \|\mathcal{D}_i(\widehat{\mathbf{S}}^{\text{body},+}) - \mathcal{D}_i(\mathbf{S}^{\text{body},+})\|_2^2, \tag{8}$$

where $\mathcal{D}_0 = \text{FK}$ gives joint poses, including sole positions at contact, while $\mathcal{D}_1 = \Delta$ and $\mathcal{D}_2 = \Delta^2$ compute first and second frame differences of body features. Temporal terms emphasize root motion and primitive boundaries. Foot support is supervised by

$$\mathcal{L}_{\text{skate}} = \|\widehat{\mathbf{v}}_f\|_{\mathcal{C}}^2 - \log p_c, \tag{9}$$

where $\widehat{\mathbf{v}}_f$ is predicted sole velocity, $\mathcal{C}$ weights recorded or predicted contact, and $p_c$ is the predicted probability of the recorded foot-contact label. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{state}} + \mathcal{L}_{\text{EE}} + \mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{skate}}. \tag{10}$$

**Inference.** For each primitive, we keep $\mathbf{S}^{\text{body},-}$ and $\mathbf{c}$ fixed, initialize the future segment with $\boldsymbol{\epsilon}$, and integrate the flow using $v_\theta^+$ from $\tau = 0$ to 1 in eight uniform Euler steps. To generate $\mathbf{S}^{\text{body}}_{t,2}$, we use the last $H$ frames of $\mathbf{S}^{\text{body}}_{t,1}$ as history and advance the EE window by $F$ frames to obtain $\mathbf{c}_{t,2}$. We then repeat the same sampling procedure in the shared local coordinate frame.

### C. Integrated Deployment System

Figure 4 shows the asynchronous planning and execution loop. To keep EE targets fixed as the robot moves, we anchor each relative DP prediction to the hand pose at its source observation:

$$\mathbf{T}_{\text{target}} = \mathbf{T}_{\text{obs}}\mathbf{T}_{\text{rel}}, \tag{11}$$

where $\mathbf{T}_{\text{obs}}$ is the hand pose at that observation and $\mathbf{T}_{\text{rel}}$ is the predicted relative EE pose. **WB-UMI** uses these anchored targets and the current whole-body history to generate a new motion reference. Once it is ready, we trim the prefix that can no longer be executed and blend the remaining reference into ongoing execution:

$$\mathbf{q}_{\text{cmd}}[j] = \text{Blend}_{w_j}(\mathbf{q}_{\text{old}}[j], \mathbf{q}_{\text{new}}[j+\delta]), \tag{12}$$

where $\delta$ is the frame offset aligning the incoming reference $\mathbf{q}_{\text{new}}$ with the active reference $\mathbf{q}_{\text{old}}$ in execution time, and the blend weight $w_j$ gradually increases from zero to one. Replanning advances by $F$ planner frames, with the second primitive serving as a preview. At each planning boundary, controller feedback confirms the active reference and playback progress. We then update the whole-body history for the next **WB-UMI** generation:

$$\mathbf{S}^{\text{body},-}_{t+1} = \mathbf{m}_{\text{fb}} \odot \mathbf{S}^{-}_{\text{meas}} + (1 - \mathbf{m}_{\text{fb}}) \odot \mathbf{S}^{-}_{\text{ref}}, \tag{13}$$

where $\mathbf{S}^{-}_{\text{meas}}$ and $\mathbf{S}^{-}_{\text{ref}}$ are measured and reference histories aligned in time and coordinates. The feedback mask $\mathbf{m}_{\text{fb}}$ selects measured channels, including joints and base tilt, and retains reference values elsewhere.

## IV. Experiments

In this section, we evaluate **WB-UMI** and the proposed hierarchy through the following questions:

- **Q1: Motion generation quality.** How well does **WB-UMI** generate coordinated whole-body motion from EE trajectories on unseen action categories?
- **Q2: Execution performance.** How accurately are the generated references tracked in simulation?
- **Q3: UMI skill transfer.** Can the proposed hierarchy enable whole-body manipulation on a physical humanoid from native UMI demonstrations, without task-specific whole-body demonstrations?

### A. Motion Generation Quality

*(1) Setup:* Using the settings in Table IV, we evaluate **WB-UMI** offline on EE trajectories extracted from mocap data, with generated motion providing the whole-body history during autoregressive rollout. The G1 dataset is built from BONES-SEED [28], retaining basic locomotion and manipulation while excluding complex motions that rarely occur in whole-body manipulation. It contains approximately 105 hours: 90 for training, 5 for validation, and 10 for testing. We split clips by semantic action group (e.g., "axe splitting wood"), with no group shared across splits, so the test categories are unseen during training.

*(2) Comparisons:* We compare **WB-UMI** with a variant without EE look-ahead and oracle variants given ground-truth root poses or root and foot poses. All variants are retrained using the same data and backbone and evaluated with the same sampling procedure. The oracle conditions cover the same history and prediction frames as the EE conditions, excluding look-ahead. These additional body signals are unavailable in native UMI demonstrations.

*(3) Metrics:* EE position, velocity, and orientation errors measure tracking of the input trajectories. Joint-rotation and FK joint-position errors measure reconstruction of the recorded whole-body motion, which is one possible realization of the EE constraints. Foot velocity during contact

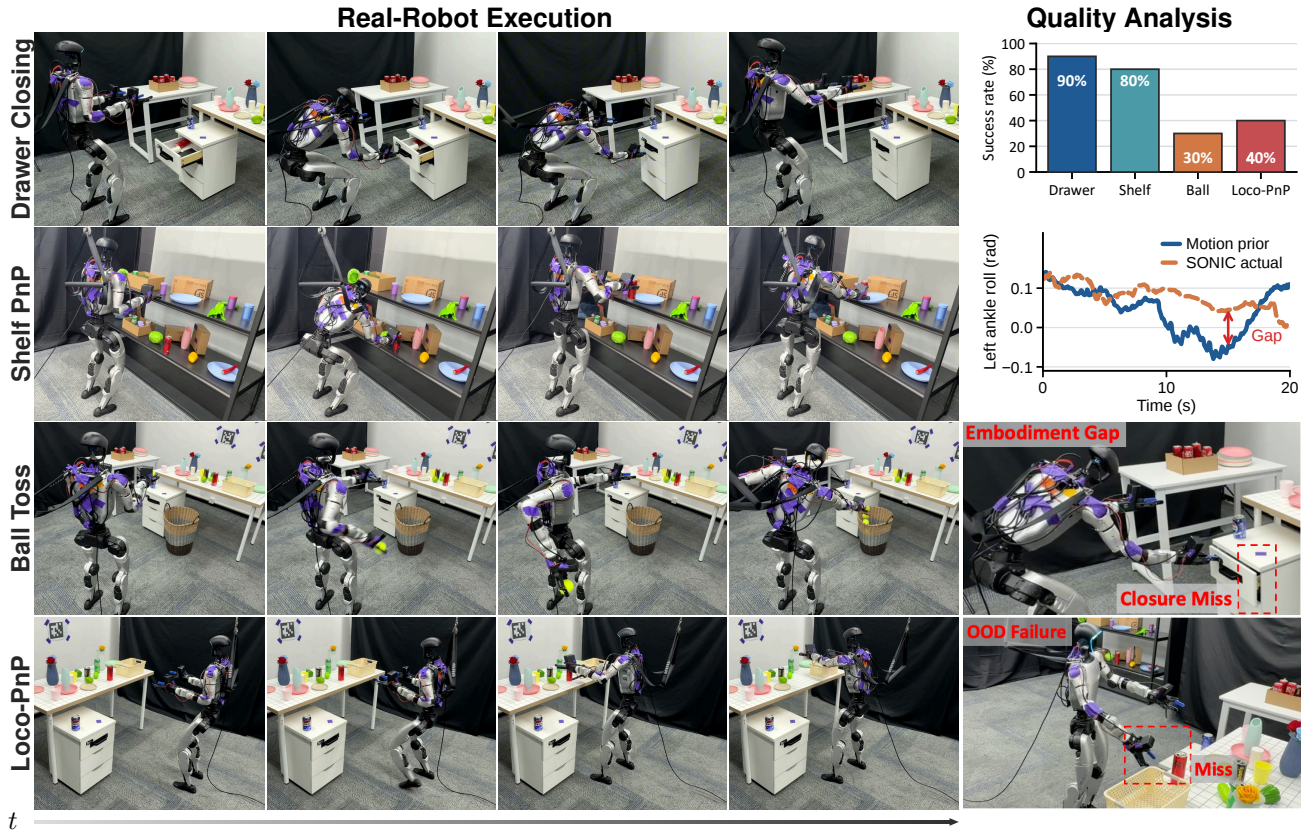


Fig. 5. Real-robot transfer with **WB-UMI**. Left: drawer closing, shelf PnP, ball toss, and Loco-PnP. Right (top to bottom): success rates, left ankle roll tracking, and drawer-closing and Loco-PnP failures. Solid blue and dashed orange curves show the **WB-UMI** reference and SONIC execution, respectively. Red annotations mark tracking errors and failures. These results demonstrate successful UMI skill transfer, with remaining limitations in reachability and whole-body tracking.

TABLE IV: Configurations of the DP and **WB-UMI**.

| **Setting** | **Value** |
|---|---|
| **DP** | |
| Input horizons (vis. / prop.) | 2 / 3 |
| Observation / action frequency | 10 / 10 Hz |
| Action horizon | 40 |
| Visual input / encoder | $2 \times 224 \times 224$; DINOv3 ViT-B/16 |
| DP backbone | DiT; 6 blocks, width 384, 6 heads |
| Learning rates (DP / vision) | $10^{-4}$ / $10^{-5}$ |
| Training | 100 epochs; batch 16; flow matching |
| **WB-UMI** | |
| History / prediction horizon | 8 / 32 |
| Rollout | 2 primitives at 30 Hz |
| EE look-ahead horizon / stride | 32 / 4 |
| Temporal backbone | DiT; 10 blocks, width 384, 6 heads |
| Learning rate | $5.66 \times 10^{-4}$ |
| Training | 300 epochs; batch 8192; masked flow matching |
| Inference | 8 ODE steps |

measures sliding. These framewise metrics are averaged within each clip and then across clips. FID measures distributional similarity between generated and recorded motions in a learned feature space. The encoder is trained only on real motions from the training split and kept fixed during evaluation. Using equally weighted clip features, we compute

$$\mathrm{FID} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|_2^2 + \mathrm{Tr}\Big(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_r - 2(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_r)^{1/2}\Big), \quad (14)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the feature mean and covariance, and $g$ and $r$ denote generated and recorded motions, respectively.

*(4) Results:* On unseen action categories, **WB-UMI** achieves an EE position error of 4.25 cm during autoregressive rollout (Table V). Figure 2(b,c) illustrates leg and trunk coordination across manipulation and locomotion, as well as changes in whole-body motion with EE trajectory speed. The oracle comparisons show different effects of additional body constraints: root and foot conditions together improve body reconstruction and reduce sliding relative to **WB-UMI**, whereas root conditions alone increase sliding. Back to **Q1**: **WB-UMI** generates coordinated whole-body motion from EE trajectories on unseen action categories, while EE look-ahead improves all reported generation metrics.

TABLE V: Offline generation on unseen action categories. Bold compares the two EE-only variants; oracle rows assess additional body conditioning.

| Condition | EE pos. (cm)↓ | EE vel. (m/s)↓ | EE orient. (deg)↓ | Foot vel. (m/s)↓ | Joint rot. (deg)↓ | Joint pos. (m)↓ | FID ↓ |
|---|---|---|---|---|---|---|---|
| **WB-UMI** w/o look-ahead | 5.77 | 0.0438 | 2.52 | 0.0590 | 5.31 | 0.0583 | 0.0280 |
| **WB-UMI** | **4.25** | **0.0364** | **1.42** | **0.0536** | **4.27** | **0.0439** | **0.0202** |
| EE+root (oracle) | 3.93 | 0.0439 | 2.43 | 0.0669 | 4.31 | 0.0289 | 0.0175 |
| EE+root+feet (oracle) | 3.87 | 0.0323 | 1.78 | 0.0236 | 2.56 | 0.0163 | 0.0079 |
| Reference motion | - | - | - | 0.008 | - | - | - |

### B. Simulated Execution Performance

*(1) Setup:* We evaluate whole-body reference execution for shelf PnP in MuJoCo using a single EE trajectory selected from UMI demonstrations collected with the GenRobot DAS Gripper [29]. Each method replays this trajectory 100 times without the DP, using identical initial states and safety limits and no external disturbances. All rollouts completed the recorded trajectory without falls or early termination.

*(2) Comparisons:* **WB-UMI** generates whole-body references for the pretrained SONIC controller. Our EE-only baseline is *Direct EE-WBC*, trained with reinforcement learning using randomly sampled EE targets as task commands. We also include IK (`Mink` [30]) as a reference under a manually specified, fixed support-contact configuration for shelf PnP. It supplies whole-body references to the same SONIC controller, but requires additional contact information beyond the EE commands, so it is not an EE-only baseline. We compare **WB-UMI** w/o feedback and **WB-UMI** (full) using the same trained checkpoint. The full variant updates its body history with measured robot state at each replanning step, while the variant without feedback uses body history from the previous motion prediction.

*(3) Metrics:* EE velocity and orientation errors measure adherence to the shared commands. Foot slip measures mean horizontal ankle-link speed during inferred stance. Root-relative mean per-joint position error (MPJPE) and joint-rotation error measure tracking of *each method's own reference*. These two metrics do not apply to Direct EE-WBC, which produces no whole-body reference. We first average each metric over frames within each rollout, then compute the mean and standard deviation across the 100 rollouts.

TABLE VI: Shelf PnP execution over 100 repetitions of a single UMI trajectory per method (mean $\pm$ standard deviation; lowest means in bold).

| Metric (↓) | IK (ref.) | Direct EE-WBC | **WB-UMI** w/o feedback | **WB-UMI** (full) |
|---|---|---|---|---|
| EE vel. (cm/s) | $\mathbf{5.40 \pm 0.50}$ | $17.49 \pm 6.50$ | $7.85 \pm 1.02$ | $8.36 \pm 1.11$ |
| EE orient. (deg) | $12.47 \pm 1.17$ | $38.89 \pm 16.41$ | $6.76 \pm 0.50$ | $\mathbf{4.97 \pm 0.47}$ |
| Foot slip (cm/s) | $2.30 \pm 1.30$ | $7.05 \pm 10.55$ | $\mathbf{0.87 \pm 0.54}$ | $0.96 \pm 0.63$ |
| Root-rel. MPJPE (cm) | $2.00 \pm 0.20$ | – | $2.36 \pm 0.20$ | $\mathbf{1.73 \pm 0.16}$ |
| Joint rot. err. (deg) | $7.28 \pm 0.46$ | – | $4.33 \pm 0.31$ | $\mathbf{2.65 \pm 0.16}$ |

*(4) Results:* On shelf PnP, both **WB-UMI** variants yield lower mean EE velocity and orientation errors and less foot slip than Direct EE-WBC (Table VI). Measured-state feedback reduces whole-body reference tracking and EE orientation errors, while slightly increasing EE velocity error and foot slip. The IK reference yields the lowest EE velocity error under a manually specified contact configuration, whereas **WB-UMI** (full) yields lower EE orientation error and foot slip without this specification. These results answer **Q2**: In the simulation, **WB-UMI** achieves lower EE velocity and orientation errors and less foot slip than Direct EE-WBC. Measured-state feedback improves whole-body reference tracking and EE orientation, with small increases in EE velocity error and foot slip.

### C. UMI Skill Transfer

We evaluate UMI skill transfer on four tasks with the 29-DoF G1 humanoid (Fig. 5), testing only the proposed hierarchy for safety. For each task, we train the DP on 200 native UMI demonstrations collected with the GenRobot DAS Gripper [29] in approximately 90 minutes. **WB-UMI** learns its whole-body motion prior independently from mocap, requiring no task-specific whole-body demonstrations. The G1 uses the corresponding DAS Controller [31] to maintain a consistent gripper interface. We conduct 10 trials per task; success requires completing the task while remaining upright and balanced.

*(1) Runtime:* Across real-robot trials on an NVIDIA GeForce RTX 4090, mean inference latencies are 64 ms for the DP and 104 ms for **WB-UMI**. Both are below their respective update intervals of 100 ms and approximately 1.07 s.

*(2) Drawer Closing:* Drawer closing tests coordinated reaching and body lowering. The robot follows the EE plan to reach the handle and fully close the drawer, succeeding in 9/10 trials. The closure miss in Fig. 5 illustrates an embodiment gap: human demonstrators can reach the target by crouching, whereas G1's shorter reach may require additional stepping or leaning.

*(3) Shelf PnP:* The robot transfers a bottle from the lower shelf to the upper shelf, coordinating arm motion with changes in body height. Success requires stable placement without dropping the bottle. The hierarchy succeeds in 8/10 trials, demonstrating whole-body manipulation across different working heights.

*(4) Ball Toss:* The robot must release a ball so that it lands in a target basket, testing dynamic whole-body coordination. The hierarchy succeeds in 3/10 trials. Failures were associated with gripper-release latency and DP prediction errors during fast EE motion.

*(5) Loco-PnP:* Loco-PnP combines locomotion and manipulation: the robot must walk to a table, grasp a bottle, and place it at the target location without dropping it. The hierarchy succeeds in 4/10 trials. The failures in Fig. 5 suggest that execution drift can expose the DP to out-of-distribution (OOD) observations, degrading subsequent EE predictions. These predictions can, in turn, increase errors in motion generation and execution. The left ankle roll curves in Fig. 5 show a gap between the generated reference and SONIC's executed motion, illustrating imperfect whole-body tracking on the real robot.

Taken together, these results answer **Q3**: the hierarchy transfers native UMI skills without task-specific whole-body demonstrations, although reliability remains lower for ball toss and Loco-PnP.

## V. Conclusion

We presented **WB-UMI**, a task-agnostic model that generates real-time whole-body motion from EE trajectories. By

learning its motion prior independently from mocap, **WB-UMI** connects a DP trained on native UMI demonstrations to humanoid execution without body trackers or task-specific whole-body demonstrations. On unseen action categories, EE look-ahead reduced the position error from 5.77 cm to 4.25 cm and improved all other reported generation metrics. The integrated system operated in real time and achieved success rates of 90%, 80%, 30%, and 40% on drawer closing, shelf pick-and-place, ball toss, and Loco-PnP, respectively. These results support reliable transfer on the first two tasks and initial feasibility on the more dynamic tasks. Dynamic feasibility and error accumulation across EE planning, motion generation, and control remain challenges. Future work will explore WBC data augmentation, and residual reinforcement learning to improve execution. Scaling pretraining to more diverse UMI and mocap datasets may further improve generalization across tasks and humanoid platforms.

## REFERENCES


[1] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[2] R. Nai, B. Zheng, J. Zhao, H. Zhu, S. Dai, Z. Chen, Y. Hu, Y. Hu, T. Zhang, C. Wen, and Y. Gao, "Humanoid manipulation interface: Humanoid whole-body manipulation from robot-free demonstrations," arXiv:2602.06643, 2026.

[3] H. Wang, C. Yu, Y. Hu, J. Zhang, Y. Li, and S. Luo, "Bifrostumi: Bridging robot-free demonstrations and humanoid whole-body manipulation," arXiv:2605.03452, 2026.

[4] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "HOMIE: Humanoid Loco-Manipulation with Isomorphic Exoskeleton Cockpit," in *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025.

[5] Y. Ze, Z. Chen, J. P. Araújo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu, "TWIST: Teleoperated whole-body imitation system," *arXiv:2505.02833*, 2025.

[6] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "Twist2: Scalable, portable, and holistic humanoid data collection system," arXiv:2511.02832, 2025.

[7] Zhaxizhuoma, K. Liu, C. Guan, Z. Jia, Z. Wu, X. Liu, T. Wang, S. Liang, P. Chen, P. Zhang, H. Song, D. Qu, D. Wang, Z. Wang, N. Cao, Y. Ding, B. Zhao, and X. Li, "Fastumi: A scalable and hardware-independent universal manipulation interface with dataset," arXiv:2409.19499, 2025.

[8] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "UMI on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," in *Proceedings of the 2024 Conference on Robot Learning*, 2024.

[9] X. Xu, J. Park, H. Zhang, E. Cousineau, A. Bhat, J. Barreiros, D. Wang, J. Bohg, and S. Song, "Hommi: Learning whole-body mobile manipulation from human demonstrations," arXiv:2603.03243, 2026.

[10] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. E. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of humanoid robots: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023, doi: 10.1109/TRO.2023.3236952.

[11] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023.

[12] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," in *The Twelfth International Conference on Learning Representations*, 2024.

[13] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[14] Y. Xie, V. Jampani, L. Zhong, D. Sun, and H. Jiang, "OmniControl: Control any joint at any time for human motion generation," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[15] E. Pinyoanuntapong, M. Saleem, K. Karunratanakul, P. Wang, H. Xue, C. Chen, C. Guo, J. Cao, J. Ren, and S. Tulyakov, "MaskControl: Spatio-Temporal Control for Masked Motion Synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 9955–9965, doi: 10.1109/ICCV51701.2025.00928.

[16] D. Rempe, M. Petrovich, Y. Yuan, H. Zhang, X. B. Peng, Y. Jiang, T. Wang, U. Iqbal, D. Minor, M. de Ruyter, J. Li, C. Tessler, E. Lim, E. Jeong, S. Wu, E. Hassani, M. Huang, J.-B. Yu, C. Chung, L. Song, O. Dionne, J. Kautz, S. Yuen, and S. Fidler, "Kimodo: Scaling controllable human motion generation," arXiv:2603.15546, 2026.

[17] K. Zhao, G. Li, and S. Tang, "DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[18] K. Zhao, M. Petrovich, H. Zhang, T. Wang, S. Tang, and D. Rempe, "Ardy: Autoregressive diffusion with hybrid representation for interactive human motion generation," *ACM Transactions on Graphics (TOG)*, vol. 45, no. 4, 2026.

[19] W. Xie, J. Zheng, J. Han, J. Shi, W. Zhang, C. Bai, and X. Li, "TextOp: Real-time interactive text-driven humanoid robot motion generation and control," arXiv:2602.07439, 2026.

[20] A. Escande, N. Mansard, and P.-B. Wieber, "Hierarchical quadratic programming: Fast online humanoid-robot motion generation," *The International Journal of Robotics Research*, vol. 33, no. 7, pp. 1006–1028, 2014, doi: 10.1177/0278364914521306.

[21] A. Herzog, N. Rotella, S. Mason, F. Grimminger, S. Schaal, and L. Righetti, "Momentum control with hierarchical inverse dynamics on a torque-controlled humanoid," *Autonomous Robots*, vol. 40, no. 3, pp. 473–491, 2016, doi: 10.1007/s10514-015-9476-6.

[22] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "DeepMimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 143:1–143:14, 2018, doi: 10.1145/3197517.3201311.

[23] Z. Zhang, J. Guo, C. Chen, J. Wang, C. Lin, Y. Lian, H. Xue, Z. Wang, M. Liu, J. Lyu, H. Liu, H. Wang, and L. Yi, "Track any motions under any disturbances," arXiv:2509.13833, 2025.

[24] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," arXiv:2508.08241, 2025.

[25] Z. Luo, Y. Yuan, T. Wang, C. Li, F. Castañeda, S. Chen, Z.-A. Cao, J. Li, D. Minor, Q. Ben, J. Park, D. Sami, Z. Wang, X. Da, R. Ding, C. Hogg, L. Song, E. Lim, E. Jeong, T. He, H. Xue, W. Xiao, S. Yuen, J. Kautz, Y. Chang, U. Iqbal, L. J. Fan, and Y. Zhu, "SONIC: Supersizing motion tracking for natural humanoid whole-body control," *Science Robotics*, vol. 11, no. 117, p. eaed4592, 2026, doi: 10.1126/scirobotics.aed4592.

[26] Y. Li, Z. Luo, T. Zhang, C. Dai, A. Kanervisto, A. Tirinzoni, H. Weng, K. Kitani, M. Guzek, A. Touati, A. Lazaric, M. Pirotta, and G. Shi, "BFM-Zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning," arXiv:2511.04131, 2025.

[27] W. Zeng, K. Yin, X. Niu, S. Lu, W. Zhong, J. Chen, F. Jia, X. Chen, Z. Wang, F. Xu, M. Zhou, K. Li, W. Zhang, H. Wang, L. Yi, D. Lin, J. Pang, and J. Wang, "Scaling behavior foundation model for humanoid robots," arXiv:2607.15163, 2026.

[28] Bones Studio, "AI Datasets for Machine Learning and Motion Capture," 2026, accessed: 2026. [Online]. Available: https://bones.studio/datasets

[29] GenRobot AI, "DAS Gripper," [Online]. Available: https://www.genrobot.ai/products/das, accessed: Aug. 24, 2026.

[30] K. Zakka, "Mink: Python inverse kinematics based on MuJoCo," Software, version 1.1.0, Feb. 2026, apache-2.0 License. [Online]. Available: https://github.com/kevinzakka/mink

[31] GenRobot AI, "DAS Controller," [Online]. Available: https://www.genrobot.ai/products/controller, accessed: Aug. 24, 2026.