VIT-H: GENERATIVE VIDEO-TO-HUMANOID CONTROL WITH MOTION TRACKING FEEDBACK

Anonymous authors
Paper under double-blind review



Figure 1: **ViT-H transfers human motions from monocular videos to a Unitree G1 humanoid.** The top rows show offline video-to-humanoid control with full-sequence generation and tracking. The bottom rows show real-time autoregressive teleoperation for locomotion and loco-manipulation.

ABSTRACT

Learning humanoid control from monocular human videos offers a scalable way to acquire diverse whole-body behaviors. Yet, existing reconstruction-retargeting-tracking pipelines are limited in system-level scalability by separately optimized intermediate stages and cannot directly optimize reference motion executability using downstream tracking feedback. We present **ViT-H**, a video-to-humanoid control framework following a generation-and-tracking paradigm. **ViT-H** directly generates humanoid reference motions from videos with a diffusion model, bypassing human motion reconstruction and retargeting, and uses motion-tracking outcomes to post-train the generator with direct preference optimization. We further introduce an autoregressive variant that enables real-time humanoid video teleoperation with sub-400 ms latency. Extensive experiments on the Unitree G1 demonstrate that **ViT-H** outperforms state-of-the-art reconstruction-retargeting-tracking pipelines in both motion quality and executability. We further show that **ViT-H** scales favorably with model size and with both pre-training and post-training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

data scale, while motion-tracking feedback consistently improves executability across different motion trackers. We will release our code and processed datasets.

1 INTRODUCTION

Controlling humanoid robots from human motions depicted in monocular videos provides a scalable way to acquire diverse whole-body behaviors from human demonstrations. Recent works have leveraged human videos to learn expressive whole-body skills (He et al., 2025; Xie et al., 2025; Sleiman et al., 2026), environment-aware behaviors that interact with surrounding geometry (Allshire et al., 2025), and interactive skills involving objects and humans (Wang et al., 2026b). Most existing systems realize this capability through a reconstruction-retargeting-tracking pipeline, which first reconstructs human motion from video, then retargets it to the humanoid embodiment, and finally tracks the resulting reference motion. This modular separation is increasingly practical as recent advances in general whole-body tracking enable a shared tracker to execute diverse reference motions (Luo et al., 2026). However, the upstream reconstruction-and-retargeting stages remain an indirect route from video to humanoid motion, where errors can accumulate to produce inconsistent, distorted, or dynamically infeasible references, making downstream tracking more difficult.

Prior work has sought to improve individual stages of this pipeline through improved motion reconstruction (Shen et al., 2024; Li et al., 2025b) and motion retargeting (Araujo et al., 2026). However, such stage-local remedies expose two structural limitations of the pipeline. First, its system-level scalability remains bounded by the remaining stages. Human motion reconstruction is not optimized for the morphology and dynamics of the target humanoid, while retargeting can only operate on reconstructed motions after upstream errors have propagated and discarded visual evidence can no longer be revisited. Second, executability is difficult to optimize end-to-end. Tracking results are observed only after reconstruction and retargeting, making it difficult to use execution outcomes to directly improve the complete video-to-humanoid mapping. Consequently, plausible reference motions at intermediate stages may still be difficult for the control policy to track.

To this end, we introduce **ViT-H**, a generative **Vi**deo-**T**o-**H**umanoid control framework following a generation-and-tracking paradigm. At the core of **ViT-H** are (i) a motion generator that maps monocular videos directly into the target humanoid motion space, eliminating the intermediate bottleneck of human motion reconstruction and retargeting, and (ii) whole-body control policies that not only track the generated reference motions but also evaluate their executability, providing motion-tracking feedback for post-training the generator.

Concretely, our motion generator is a video-conditioned humanoid motion diffusion model first pre-trained on paired videos and motions. The whole-body control policies are fine-tuned from a pre-existing general motion tracker, yielding a deployable tracking policy and a privileged variant with additional global observations. We evaluate motions sampled from the generator with both tracking policies and use the resulting tracking outcomes to construct preference pairs for direct preference optimization (DPO), explicitly incorporating executability into the humanoid motion generator. Additionally, we extend **ViT-H** with an autoregressive variant of the motion generator that produces motions in temporal segments, enabling online control from streaming video and real-time humanoid video teleoperation.

Through extensive experiments, we demonstrate that our motion generator produces humanoid motions with improved motion quality and executability over state-of-the-art reconstruction-and-retargeting baselines. Our ablation studies further show that DPO post-training delivers substantial gains in motion executability over the corresponding pre-training-only generators. Additionally, we demonstrate that **ViT-H** scales favorably with model capacity and both pre-training and post-training data size, while its post-training gains transfer across multiple motion trackers, indicating improved general executability rather than overfitting to a particular motion tracker.

Our primary contributions are threefold. First, we propose **ViT-H**, a generative video-to-humanoid control framework that directly generates executable humanoid motions from monocular videos without human motion reconstruction and retargeting. Second, we introduce post-training with motion-tracking feedback, which explicitly incorporates executability into reference motion generation via DPO. Third, extensive experiments demonstrate that **ViT-H** outperforms state-of-the-art reconstruction-retargeting-tracking baselines in motion quality and executability, scales favorably

with model capacity and pre-training and post-training data size, and transfers its post-training gains across multiple motion trackers.

2 Related Work

Human motion reconstruction from video. Monocular 3D human reconstruction has been extensively studied with parametric body models (Loper et al., 2015; Pavlakos et al., 2019; Saito et al., 2026). Early approaches primarily estimate human pose and shape from individual images (Kanazawa et al., 2018; Kolotouros et al., 2019; Kocabas et al., 2021; Li et al., 2021; Zhang et al., 2023), while video-based methods exploit temporal information to improve motion consistency and robustness (Kocabas et al., 2020; Wan et al., 2021; Shen et al., 2023). Recent work extends human motion recovery beyond camera coordinates by estimating globally consistent trajectories and disentangling human and camera motion (Yuan et al., 2022; Ye et al., 2023; Sun et al., 2023; Kocabas et al., 2024; Shin et al., 2024; Wang et al., 2024; Shen et al., 2024; Li et al., 2025b; Zhao et al., 2026b). Complementary approaches improve the physical plausibility of reconstructed human motions through dynamics-aware refinement, motion restoration, or physics-based character control (Zhang et al., 2024; 2025; Feng et al., 2025). While these methods operate in the human motion space, our goal is not to recover physically plausible human motion as an intermediate representation, but to directly generate robot-space humanoid reference motions and optimize them for downstream tracking executability.

Human-to-humanoid motion retargeting. Motion retargeting converts human motions into humanoid robot joint space through geometric alignment and optimization-based inverse kinematics (He et al., 2024b;a). GMR (Araujo et al., 2026) improves general-purpose retargeting by reducing kinematic artifacts and preserving source-motion fidelity, while PHUMA (Lee et al., 2026) introduces physical constraints to enforce joint limits and contact consistency. Concurrent work NMR (Zhao et al., 2026a) learns a direct human-to-humanoid motion mapping to facilitate downstream whole-body tracking. In contrast, our approach bypasses human-to-humanoid retargeting altogether and directly generates robot-specific reference motions from monocular videos.

Humanoid motion generation and whole-body tracking. Whole-body motion tracking provides a scalable interface for executing diverse kinematic reference motions on humanoid robots (Ze et al., 2025; Chen et al., 2025; Liao et al., 2026; Luo et al., 2026; Qi et al., 2026; Zeng et al., 2026; Chen et al., 2026b). Building on such trackers, recent works generate text-conditioned humanoid motion references for whole-body control (Jiang et al., 2024c; Rempe et al., 2026; Zhang et al., 2026b; Harithas et al., 2026), extend motion generation to multimodal conditioning (Huang et al., 2026), and predict humanoid kinematic trajectories for loco-manipulation (Wang et al., 2026a). Our approach also follows this generation-and-tracking paradigm, but rather than optimizing the two components independently, we use tracking feedback to post-train the motion generator for executability.

Another line of work further collapses explicit motion generation and tracking, enabling retargeting-free humanoid control from language (Li et al., 2026e), audio (Li et al., 2026c), or video (Li et al., 2026d) by training dedicated control policies directly conditioned on motion latents. While this formulation eliminates explicit motion decoding, it requires a specialized control policy for the latent interface and is therefore less able to directly leverage advances in general-purpose motion trackers.

Post-training of human motion generation models. Recent works post-train motion generation models with preference or reward signals to improve motion quality and alignment (Pappa et al., 2026; Tan et al., 2025). Another line of work incorporates downstream physical feedback into text-conditioned human motion generation (Yue et al., 2025; Zhang et al., 2026a), using physics-based tracking or a whole-body controller to improve motion plausibility. Despite enabling deployment on humanoid robots, these methods require motion retargeting and optimize motions in the human motion space rather than for the target humanoid embodiment. Our approach instead post-trains the video-conditioned motion generator directly in the target humanoid motion space.

3 ViT-H

Formulation. Our goal is to convert human motions depicted in monocular videos $\{\mathbf{I}_t\}_{t=1}^{T}$ into executable whole-body actions that enable a Unitree G1 humanoid to reproduce the demonstrated

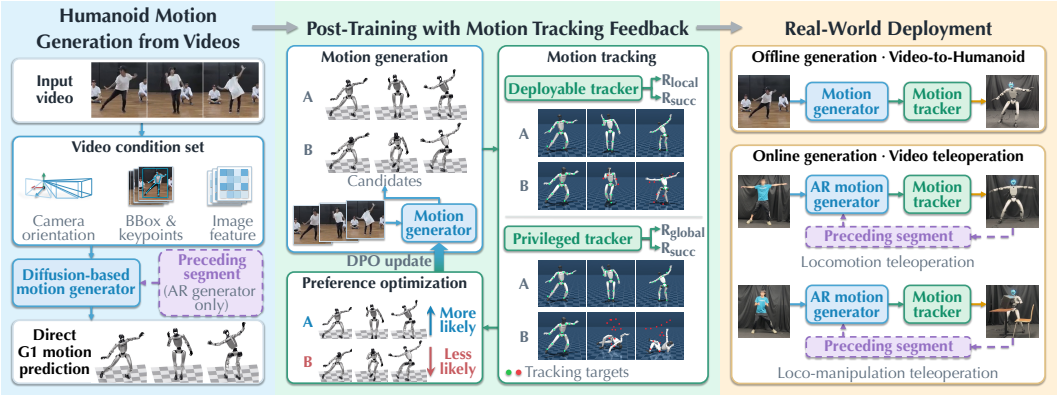


Figure 2: **Overview of ViT-H. Left:** We train a diffusion-based motion generator to directly predict G1 reference motions from video. We further train an autoregressive variant with preceding-segment conditioning. **Middle:** Both generators are post-trained with DPO, using our deployable motion tracker and its privileged variant to provide tracking-based rewards. **Right:** At deployment, **ViT-H** enables video-to-humanoid control with the offline generator and supports real-time video teleoperation with the autoregressive online generator.

motions. We formulate this video-to-action control problem as reference motion generation followed by whole-body tracking: a video-conditioned generative model π_{gen} predicts a reference humanoid motion $\mathbf{X} \sim \pi_{\text{gen}}(\cdot|\mathbf{I})$, and a whole-body tracking policy π_{track} produces actions $\mathbf{a}_t \sim \pi_{\text{track}}(\cdot \mid \mathbf{s}_t, \mathbf{X}_t)$ to execute the generated motion with a proportional-derivative (PD) controller. We consider both *offline* generation, where the complete reference motion is generated at once, and *online* generation, where it is generated autoregressively in temporal segments conditioned on preceding generation.

Overview. We propose **ViT-H**, a generation-and-tracking framework for video-driven humanoid control. An overview of **ViT-H** is shown in Figure 2. We first train a diffusion-based motion generator π_{gen} to directly predict reference humanoid motions from video, eliminating the need for human motion reconstruction and retargeting (Section 3.1). We then fine-tune a pre-existing general motion tracker (Luo et al., 2026) to obtain a deployable tracking policy π_{track}, and a privileged variant $\pi_{\text{track}}^{\text{priv}}$ with access to additional global observations. Beyond executing generated reference motions, these trackers also provide executability feedback for post-training: motions sampled from π_{gen} are evaluated through tracking, and the resulting rewards are used to construct preference pairs for DPO (Rafailov et al., 2023) fine-tuning of π_{gen} (Section 3.2). For the online setting, we introduce an autoregressive variant of π_{gen} to support real-time humanoid video teleoperation (Section 3.3). We leave the implementation details for real-world deployment in Section B.7.

3.1 Humanoid Motion Generation from Videos

Data representation. We represent video frames with condition set $\mathcal{C} = \{\mathbf{c}_{\text{cam}}, \mathbf{c}_{\text{bbox}}, \mathbf{c}_{\text{kp2d}}, \mathbf{c}_{\text{feat}}\}$, where $\mathbf{c}_{\text{cam}} \in \mathbb{R}^{T\times 6}$ denotes relative camera orientations, $\mathbf{c}_{\text{bbox}} \in \mathbb{R}^{T\times 3}$ denotes 2D bounding boxes, $\mathbf{c}_{\text{kp2d}} \in \mathbb{R}^{T\times 2K}$ denotes 2D keypoints, and $\mathbf{c}_{\text{feat}} \in \mathbb{R}^{T\times 1024}$ denotes image features.

Inspired by the gravity-view coordinate system (Shen et al., 2024) introduced for human motion reconstruction, we represent the humanoid motion $\mathbf{X}$ with: (i) root velocity $\mathbf{v}_{\text{root}} \in \mathbb{R}^{T\times 3}$, camera-view root orientation $\mathbf{\Gamma}_{\text{cam}} \in \mathbb{R}^{T\times 6}$, and gravity-view root orientation $\mathbf{\Gamma}_{\text{gv}} \in \mathbb{R}^{T\times 6}$ to parameterize the global trajectory, and (ii) local joint rotations $\mathbf{Q} \in \mathbb{R}^{T\times J\times 6}$ encoded using the 6D representation (Zhou et al., 2019). Additionally, we predict the camera-relative scale and translation $\boldsymbol{\tau} \in \mathbb{R}^{T\times 3}$ and end-effector stationary probabilities $\mathbf{p}_{\text{ee}} \in \mathbb{R}^{T\times 6}$, resulting in the full humanoid motion $\mathbf{X} = (\mathbf{v}_{\text{root}}, \mathbf{\Gamma}_{\text{cam}}, \mathbf{\Gamma}_{\text{gv}}, \mathbf{Q}, \boldsymbol{\tau}, \mathbf{p}_{\text{ee}})$. Details on data representation are in Section B.1.

Video-conditioned humanoid motion generation. The reference humanoid motion generator π_{gen} is a diffusion model that takes the video-derived condition set $\mathcal{C}$ as conditional input. Following the DDPM framework (Ho et al., 2020), we implement a forward diffusion process that progressively

corrupts clean humanoid motions with Gaussian noise over N steps:

$$q(\mathbf{X}_n|\mathbf{X}_{n-1}) = \mathcal{N}(\mathbf{X}_n; \sqrt{1-\beta_n}\mathbf{X}_{n-1}, \beta_n \boldsymbol{I}), \tag{1}$$

where n denotes the diffusion step and β_n specifies the variance schedule. The generator π_{gen} learns the reverse process, iteratively denoising samples initialized from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. We adopt the $\mathbf{X}_0$-prediction parameterization and train π_{gen} to recover the clean motion from noisy samples directly:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{n\sim\mathcal{U}\{1,\dots,N\},\mathbf{X}_n\sim q(\mathbf{X}_n|\mathbf{X}_0)} \left\| \mathbf{X}_0 - \pi_{\text{gen}}(\mathbf{X}_n, n, \mathcal{C}) \right\|_2^2. \tag{2}$$

Additionally, we recover the global trajectory from the predicted $\hat{\mathbf{X}}_0$ and apply forward kinematics to obtain joint and vertex positions, which are used to compute joint, vertex, translation, and end-effector stationary losses. Details on network architecture and losses are in Sections B.2 and B.3, respectively.

3.2 Post-Training with Motion Tracking Feedback

Humanoid motion tracking. Our whole-body tracking policy π_{track} is fine-tuned from SONIC (Luo et al., 2026), a general humanoid motion tracker trained with proximal policy optimization (PPO) (Schulman et al., 2017). We introduce two modifications during fine-tuning: (i) additional rewards for tracking the anchor position on the horizontal plane and the anchor yaw orientation, and (ii) additional critic observations containing the current and future root transformations of both the robot and the reference motion.

To assess the global trajectory predicted by π_{gen} during post-training, we additionally train a privileged whole-body tracking policy $\pi_{\text{track}}^{\text{priv}}$ with access to global observations. Specifically, $\pi_{\text{track}}^{\text{priv}}$ is fine-tuned from π_{track} by augmenting the actor observations with the relative root transformation between the robot and the reference motion, while keeping the remaining training setup unchanged. Details on whole-body tracking policy training are in Section B.6.

Preference pair construction. For each video-derived condition set $\mathcal{C}$, we sample M candidates $\{\mathbf{X}^{(i)}\}_{i=1}^{M} \sim \pi_{\text{gen}}(\cdot \mid \mathcal{C})$ using our motion generator. We denote by $\bar{\mathbf{X}}^{(i)}$ and $\bar{\mathbf{X}}_{\text{priv}}^{(i)}$ the corresponding rollout motions obtained with π_{track} and $\pi_{\text{track}}^{\text{priv}}$, respectively. We construct preference pairs using three types of rewards: a local pose reward, a global motion reward, and a tracking-success reward. Let $\mathcal{E}_{\text{root-pos}}$, $\mathcal{E}_{\text{root-ori}}$, $\mathcal{E}_{\text{pos}}$, and $\mathcal{E}_{\text{ori}}$ denote the root translation, root orientation, root-relative joint position, and joint rotation errors, respectively. We first convert each error into a normalized reward

$$\rho_k(\mathbf{X}, \mathbf{X}') = \exp\big(-\mathcal{E}_k(\mathbf{X}, \mathbf{X}')/s_k^2\big) \in (0, 1], \tag{3}$$

where s_k denotes the scale for error type k. The local and global rewards are then defined as

$$\mathcal{R}_{\text{local}}(\mathbf{X}, \mathbf{X}') = \boldsymbol{\alpha}_{\text{local}}^{\top} \big[\rho_{\text{pos}}(\mathbf{X}, \mathbf{X}'), \rho_{\text{ori}}(\mathbf{X}, \mathbf{X}')\big]^{\top}, \tag{4}$$

$$\mathcal{R}_{\text{global}}(\mathbf{X}, \mathbf{X}') = \boldsymbol{\alpha}_{\text{global}}^{\top} \big[\rho_{\text{root-pos}}(\mathbf{X}, \mathbf{X}'), \rho_{\text{root-ori}}(\mathbf{X}, \mathbf{X}'), \rho_{\text{pos}}(\mathbf{X}, \mathbf{X}'), \rho_{\text{ori}}(\mathbf{X}, \mathbf{X}')\big]^{\top}, \tag{5}$$

where $\boldsymbol{\alpha}_{\text{local}}$ and $\boldsymbol{\alpha}_{\text{global}}$ denote the normalized weights. We define the tracking-success reward as

$$\mathcal{R}_{\text{succ}}(s, s_{\text{priv}}) = \mathbb{I}[s = 1 \wedge s_{\text{priv}} = 1], \tag{6}$$

where s and s_{priv} indicate whether the corresponding trackers successfully complete the rollout.

Using these formulations, we consider three rewards that do not require ground-truth motion:

$$r_1^{(i)} = \mathcal{R}_{\text{local}}\big(\mathbf{X}^{(i)}, \bar{\mathbf{X}}^{(i)}\big), \quad r_2^{(i)} = \mathcal{R}_{\text{global}}\big(\mathbf{X}^{(i)}, \bar{\mathbf{X}}_{\text{priv}}^{(i)}\big), \quad r_3^{(i)} = \mathcal{R}_{\text{succ}}\big(s^{(i)}, s_{\text{priv}}^{(i)}\big). \tag{7}$$

When ground-truth humanoid motion $\mathbf{X}_{\text{GT}}$ is available, we additionally consider

$$r_4^{(i)} = \mathcal{R}_{\text{global}}\big(\mathbf{X}^{(i)}, \mathbf{X}_{\text{GT}}\big), \quad r_5^{(i)} = \mathcal{R}_{\text{local}}\big(\mathbf{X}_{\text{GT}}, \bar{\mathbf{X}}^{(i)}\big), \quad r_6^{(i)} = \mathcal{R}_{\text{global}}\big(\mathbf{X}_{\text{GT}}, \bar{\mathbf{X}}_{\text{priv}}^{(i)}\big). \tag{8}$$

We form a preference pair $(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$ when the reward advantage of $\mathbf{X}^{(i)}$ over $\mathbf{X}^{(j)}$ exceeds a margin $\boldsymbol{\delta}$ along each reward dimension: $\mathbf{r}^{(i)} - \mathbf{r}^{(j)} \succ \boldsymbol{\delta}$. Additionally, a sample is eligible to serve as the preferred sample only if it is successfully tracked by both trackers. Let $\mathcal{P}$ denote the set of all candidate pairs satisfying this condition. We select the pair with the largest total reward gap:

$$(i^\star, j^\star) = \arg\max_{(i,j)\in\mathcal{P}} \mathbf{w}^{\top}\big(\mathbf{r}^{(i)} - \mathbf{r}^{(j)}\big), \tag{9}$$

where $\mathbf{w}$ denotes the reward weights. We finally construct the pair $(\mathbf{X}^+, \mathbf{X}^-; \mathcal{C}) = (\mathbf{X}^{(i^\star)}, \mathbf{X}^{(j^\star)}; \mathcal{C})$.

DPO post-training. During post-training, we fine-tune π_{gen} while keeping a frozen copy of its pre-trained parameters, denoted by $\pi_{\text{gen}}^{\text{ref}}$, as the reference model. For each pair, we sample a diffusion step n and shared Gaussian noise $\boldsymbol{\epsilon}$, and obtain $\mathbf{X}_n^+ \sim q(\mathbf{X}_n^+ \mid \mathbf{X}^+)$ and $\mathbf{X}_n^- \sim q(\mathbf{X}_n^- \mid \mathbf{X}^-)$ using the same $\boldsymbol{\epsilon}$. Since our model adopts $\mathbf{X}_0$ prediction, we define the per-sample denoising loss as

$$\ell_\pi(\mathbf{X}, \mathbf{X}_n, n, \mathcal{C}) = \left\|\mathbf{X} - \pi(\mathbf{X}_n, n, \mathcal{C})\right\|_2^2. \tag{10}$$

Following Diffusion-DPO (Wallace et al., 2024), we formulate the DPO objective as:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} = -\mathbb{E}\Big[\log\sigma\Big(&\beta_{\text{DPO}}\left[\ell_{\pi_{\text{gen}}}(\mathbf{X}^-, \mathbf{X}_n^-, n, \mathcal{C}) - \ell_{\pi_{\text{gen}}^{\text{ref}}}(\mathbf{X}^-, \mathbf{X}_n^-, n, \mathcal{C})\right] \\ &-\beta_{\text{DPO}}\left[\ell_{\pi_{\text{gen}}}(\mathbf{X}^+, \mathbf{X}_n^+, n, \mathcal{C}) - \ell_{\pi_{\text{gen}}^{\text{ref}}}(\mathbf{X}^+, \mathbf{X}_n^+, n, \mathcal{C})\right]\Big)\Big],\end{aligned} \tag{11}$$

where σ denotes the sigmoid function and β_{DPO} controls the strength of optimization. We further apply supervised fine-tuning (SFT) on the preferred motion and optimize the combined objective:

$$\mathcal{L}_{\text{post-train}} = \lambda_{\text{DPO}}\mathcal{L}_{\text{DPO}} + \lambda_{\text{SFT}}\mathcal{L}_{\text{SFT}}. \tag{12}$$

Details on preference pair construction and post-training are in Section B.4.

3.3 Autoregressive Generation for Video Teleoperation

For the online setting, we extend **ViT-H** with an autoregressive motion generator $\pi_{\text{gen}}^{\text{ar}}$ that enables segment-by-segment motion generation for real-time video-to-humanoid teleoperation. Each segment has a fixed length of T frames, and consecutive segments share L overlapping frames. Each motion segment $\mathbf{X}^s$ is conditioned on the preceding segment $\mathbf{X}^{s-1}$. Following prior work (Jiang et al., 2024b;a), we inject the overlapping context by replacing the first L frames of the noisy input $\mathbf{X}_n^s$ with the last L frames of $\mathbf{X}^{s-1}$. To preserve the ability to generate the initial segment without preceding context, we randomly drop $\mathbf{X}^{s-1}$ during fine-tuning.

During post-training, $\mathbf{X}^{s-1}$ is provided as conditioning only for in-distribution samples with ground-truth motion. For video-only samples, where the preceding ground-truth motion is unavailable, we follow the same post-training procedure as in Section 3.2 without autoregressive conditioning. Details on the adaptation of the autoregressive generator are in Section B.5.

4 Experiments

In this section, we first compare the motion generator π_{gen} of **ViT-H** against reconstruction-and-retargeting baselines and our pre-training-only ablation variants on both datasets with 3D ground-truth annotations and video-only datasets (Section 4.1). We then provide an empirical analysis of DPO post-training, investigating how preference construction and post-training design choices affect humanoid motion executability (Section 4.2). We finally discuss the scalability of **ViT-H** (Section 4.3).

4.1 Comparisons

Datasets. We train π_{gen} on a mixture of AMASS (Mahmood et al., 2019), BONES-SEED (Studio, 2026), BEDLAM (Black et al., 2023), BEDLAM2.0 (Tesch et al., 2026), and Human3.6M (Ionescu et al., 2013). Datasets containing only human motions are retargeted to Unitree G1 motions using our adapted GMR retargeter (Araujo et al., 2026). During post-training, we fine-tune π_{gen} on BEDLAM, BEDLAM2.0, Human3.6M, and video-only datasets (Li et al., 2020; 2025a; 2026a) covering diverse motion categories, including dancing, fitness, gymnastics, and martial arts. Both whole-body tracking policies, π_{track} and $\pi_{\text{track}}^{\text{priv}}$, are fine-tuned exclusively on BONES-SEED.

For evaluation, we use EMDB (Kaufmann et al., 2023) and RICH (Huang et al., 2022), whose 3D ground-truth human motions are retargeted to Unitree G1 motions and filtered accordingly. We further conduct evaluation on held-out splits of the video-only datasets. Details on dataset statistics, human-to-humanoid conversion, and conversions across human body models are in Section C.1.

Evaluation metrics. For datasets with retargeted 3D ground truth, we report local MPJPE, PA-MPJPE, and root translation error (RTE). We also report motion jitter (Jitter), and self-penetration

Table 1: **Quantitative comparison of motion quality.** † denotes adapted baselines trained on our datasets with comparable model sizes. Lower is better for all quantitative metrics. For the 2AFC study, we report the preference rate over Ours (offline).

Method	EMDB					RICH					Video-only		
	MPJPE	PA-MPJPE	RTE	Jitter	Pen.	MPJPE	PA-MPJPE	RTE	Jitter	Pen.	Jitter	Pen.	vs. Ours (offline)
GVHMR + GMR	73.3	54.4	3.1	18.4	2.3	51.6	44.9	4.6	7.7	1.7	15.5	1.8	17.5
GVHMR† + GMR	71.4	49.1	3.5	16.5	1.6	69.6	51.5	4.4	10.3	2.2	16.8	2.1	23.8
GEM + GMR	62.3	38.4	4.6	19.6	1.3	56.5	49.4	4.7	11.9	2.2	23.9	2.2	16.3
GEM† + GMR	72.9	51.0	4.5	21.7	2.0	70.3	51.2	4.4	20.5	2.2	25.9	1.9	17.5
Ours (online) w/o PT	55.4	36.6	8.3	25.3	0.2	44.5	34.5	6.8	24.4	0.3	24.8	1.3	22.5
Ours (online)	55.5	36.8	6.4	18.5	**0.1**	44.3	36.1	6.3	17.4	0.2	13.6	1.2	28.8
Ours (offline) w/o PT	**52.9**	**34.0**	2.6	11.2	0.2	38.7	**28.5**	4.4	**7.2**	0.2	10.6	1.0	41.3
Ours (offline)	53.3	34.3	**2.5**	**10.6**	**0.1**	**38.4**	29.4	**4.3**	**7.2**	**0.1**	**9.5**	**0.9**	-

Table 2: **Quantitative comparison of motion executability.** † denotes adapted baselines trained on our datasets with comparable model sizes. Higher is better for Succ.; lower is better for other metrics.

Method	EMDB					RICH					Video-only		
	Succ.	MPJPE	PA-MPJPE	Trk. MPJPE	Trk. PA-MPJPE	Succ.	MPJPE	PA-MPJPE	Trk. MPJPE	Trk. PA-MPJPE	Succ.	Trk. MPJPE	Trk. PA-MPJPE
GT + GMR	91.0	35.2	22.7	35.2	22.7	100.0	30.3	19.6	30.3	19.6	-	-	-
GVHMR + GMR	61.7	101.2	59.8	47.5	29.8	90.3	64.6	47.2	36.4	28.9	50.3	91.9	52.2
GVHMR† + GMR	71.4	94.5	57.0	44.7	30.0	84.7	90.8	58.2	53.4	34.7	48.7	105.8	60.3
GEM + GMR	73.0	94.9	43.0	43.2	25.4	90.3	71.7	51.0	37.9	29.7	51.7	99.8	57.2
GEM† + GMR	73.0	100.0	58.1	43.9	29.4	90.3	90.9	55.1	52.5	34.2	46.7	106.8	59.8
GVHMR	56.4	91.3	45.7	-	-	29.2	83.5	46.4	-	-	20.5	-	-
GVHMR†	54.9	85.8	45.7	-	-	26.4	82.0	46.5	-	-	21.9	-	-
GEM	54.9	94.3	45.8	-	-	30.6	86.8	47.5	-	-	17.3	-	-
GEM†	55.6	90.1	45.8	-	-	23.6	83.7	45.9	-	-	19.5	-	-
Ours (online) w/o PT	76.9	85.8	41.3	34.0	23.1	87.5	76.0	40.9	34.3	21.6	49.6	71.7	41.9
Ours (online)	88.5	84.0	40.7	**29.4**	**16.4**	98.2	69.8	40.0	27.6	**17.3**	63.8	65.2	37.8
Ours (offline) w/o PT	89.5	75.4	40.0	32.0	18.6	**100.0**	**62.0**	**34.3**	26.3	18.2	74.8	59.0	37.1
Ours (offline)	**90.2**	**75.1**	**39.7**	31.5	18.0	**100.0**	**62.0**	34.6	**25.9**	17.7	**80.0**	**54.2**	**34.7**

(Pen.) to further evaluate motion quality. For video-only datasets, we report Jitter and Pen., and conduct a two-alternative forced-choice (2AFC) study comparing the offline generation mode of π_{gen} against the baselines and ablation variants. Details on human study are in Section C.2.

We further evaluate the generation-and-tracking pipeline of **ViT-H** by executing the predicted humanoid motions in MuJoCo with π_{track}. We report the tracking success rate (Succ.) for all evaluation datasets. For datasets with retargeted ground-truth humanoid motions, we report MPJPE and PA-MPJPE against the ground truth. Additionally, we report Trk. MPJPE and Trk. PA-MPJPE, defined as the MPJPE and PA-MPJPE between the predicted motions and their corresponding tracked motions, respectively, to measure how easily the motions can be tracked. Details on metrics are in Section C.3.

Baselines and ablation variants. We compare our video-conditioned motion generator π_{gen} against reconstruction-and-retargeting baselines. We adapt GVHMR (Shen et al., 2024) and GEM (Li et al., 2025b) as human motion reconstruction methods, followed by GMR (Araujo et al., 2026) for motion retargeting. For both reconstruction methods, we additionally train variants on our human-motion training set and adopt model sizes comparable to π_{gen}. As ablations, we evaluate the offline generator π_{gen} and its autoregressive variant $\pi^{\text{ar}}_{\text{gen}}$, both without post-training, denoted as "Ours w/o PT".

For the executability evaluation in MuJoCo, since π_{track} follows SONIC (Luo et al., 2026) in accepting human motions directly as tracking references, we evaluate the motion reconstruction baselines without retargeting. We additionally report tracking performance on the retargeted ground-truth humanoid motions as a reference. Details on baseline adaptation are in Section C.4.

Results. We report quantitative results on humanoid reference motion quality using estimated camera motion in Table 1. Specifically, we use DPVO (Teed et al., 2023) for GVHMR and π_{gen}, and Droid-SLAM (Teed & Deng, 2021) for GEM. Our offline motion generator consistently outperforms the reconstruction-and-retargeting baselines across motion accuracy and quality metrics, showing the

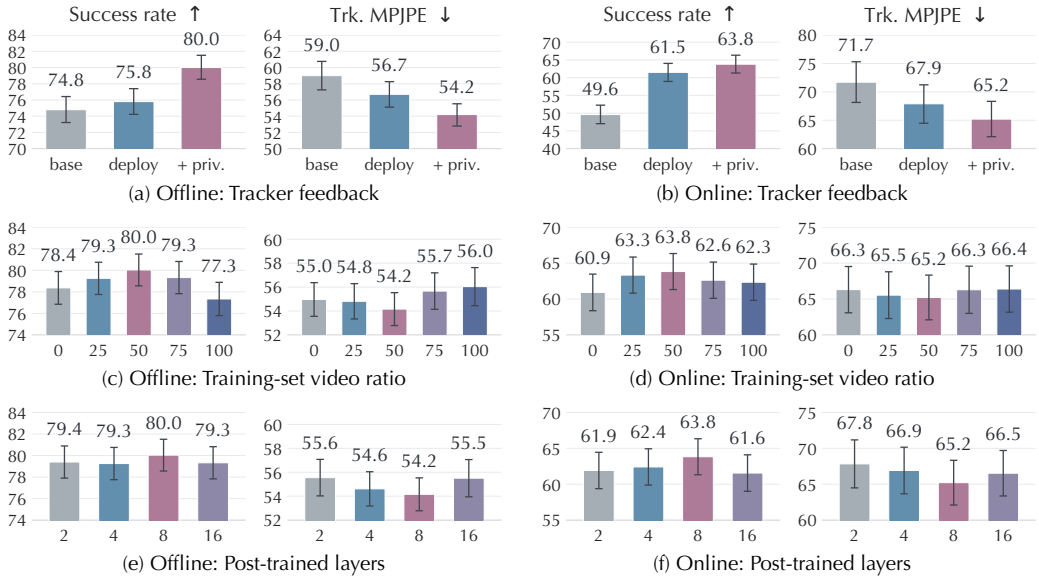


Figure 3: **Analysis of DPO post-training on the video-only out-of-distribution test set.** We study the effects of privileged tracking rewards, post-training data composition, and fine-tuning depth, and report tracking success rate and Trk. MPJPE for both generators.

advantage of directly predicting motions in the G1 embodiment. The human study on video-only datasets further shows that participants consistently prefer our results over the baselines, aligning well with the quantitative metrics. Additionally, although the online generator is less competitive on root translation accuracy under the causal setting, it still achieves lower local MPJPE than the baselines, indicating that accurate local motion can be preserved in the causal setting. Finally, comparisons with the w/o PT variants show that post-training improves motion quality, as reflected by lower motion jitter and self-penetration in most settings, while maintaining comparable motion prediction accuracy.

The quantitative results on humanoid reference motion executability are reported in Table 2, also using estimated camera motion. Both generators show clear overall improvements over the baselines, with the offline generator π_{gen} outperforming them consistently and the online generator $\pi^{\text{ar}}_{\text{gen}}$ remaining competitive across metrics. Moreover, post-training consistently improves the executability of both generators across all three test sets, with particularly large gains on the challenging video-only datasets, demonstrating the effectiveness of explicitly incorporating motion-tracking feedback into reference motion generation.

To further assess the upper-bound performance of our approach, we repeat the evaluation using ground-truth camera in Section D.1. The resulting trends are consistent with our main experiments. We leave the qualitative comparisons on the *project page* included in the *supplementary materials*.

4.2 Analysis of DPO Post-Training

We analyze three aspects of post-training: reward design, post-training data sources, and fine-tuning depth. Since EMDB and RICH are relatively less challenging under our evaluation protocol (Table 2), we focus this analysis on the more challenging video-only test set. Results are shown in Figure 3.

First, we ablate the global-motion reward computed by the privileged tracker $\pi^{\text{priv}}_{\text{track}}$. Although all evaluation metrics are computed using the deployable tracker π_{track}, incorporating the privileged reward during preference construction consistently improves performance, indicating that global tracking feedback provides complementary supervision during post-training. Second, we vary the amount of in-distribution paired data relative to the video-only out-of-distribution data. Performance exhibits a non-monotonic trend, with a moderate amount of paired data performing best empirically; increasing the paired-data proportion beyond 50% provides no further gains and can degrade performance. Finally, we vary the number of Transformer layers updated during DPO post-training. Among the 16 layers of our generator, fine-tuning the last 8 yields the best empirical performance, while updating either fewer layers or all 16 layers results in lower performance.

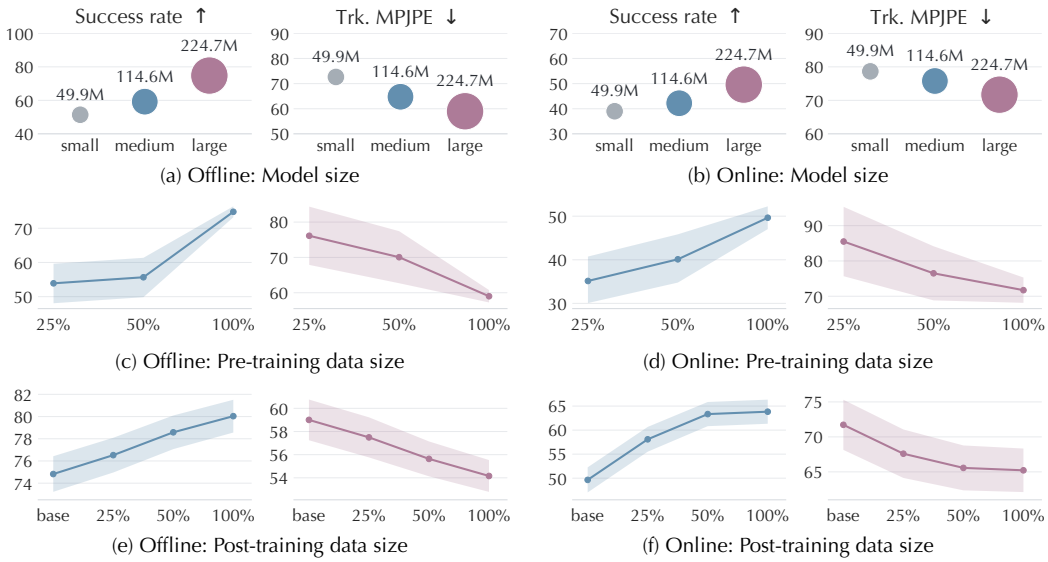


Figure 4: **Scaling behavior of ViT-H with increasing model size and training data.**

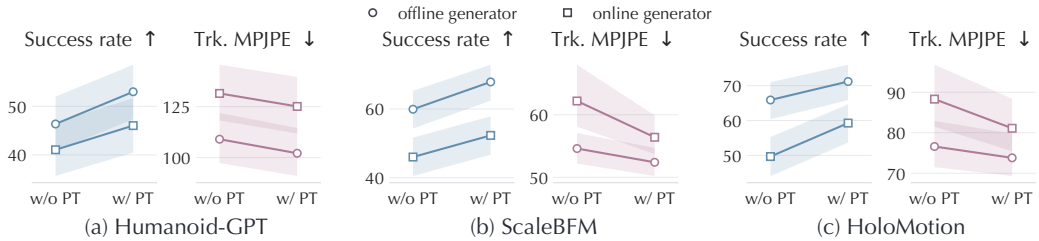


Figure 5: **Cross-tracker improvements before and after post-training.**

4.3 Scalability of ViT-H

We study how **ViT-H** scales with model capacity and training data on the video-only test set (Figure 4). Model-size and pre-training-data results are obtained without post-training. Increasing the generator from 49.9M to 114.6M and 224.7M parameters consistently raises the tracking success rate and lowers Trk. MPJPE for both the offline and the online generator. Pre-training on 25%, 50%, and 100% of the pre-training data shows the same trend, with the largest gain in success rate from 50% to 100%. Post-training data scale behaves similarly. Increasing post-training data consistently improves both metrics for both the offline and the online generators.

In Figure 5, we execute the motions generated by both generators on the video-only test set with Humanoid-GPT (Qi et al., 2026), ScaleBFM (Zeng et al., 2026), and HoloMotion (Chen et al., 2026a), none of which is used during post-training, and observe consistent improvements in both metrics after post-training across all three trackers. These results suggest that post-training improves the intrinsic executability of the generated motions rather than overfitting to a specific tracker.

5 Conclusion

We presented **ViT-H**, a video-to-humanoid control framework that maps human motions depicted in monocular videos directly to executable whole-body humanoid motions. By replacing reconstruction and retargeting with direct generation in the target humanoid motion space, and further post-training the generator with whole-body motion-tracking feedback, **ViT-H** enables scalable video-conditioned reference motion generation while directly incorporating downstream executability into the generation process. Extensive experiments show that **ViT-H** improves both motion quality and executability over baselines, scales favorably with model capacity and pre-training and post-training data size, and achieves consistent post-training gains across multiple motion trackers.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods under the authors' supervision. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Our work does not introduce synthetic datasets or contain mathematical claims or proofs, and the corresponding required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools to edit the paper to improve readability, source and search for information, and identify relevant literature. We have reviewed all AI-assisted work. Specifically, all AI-assisted method implementations were reviewed, tested, and verified for correctness by the co-first authors. We also carefully reviewed all AI-assisted edits to the manuscript and manually verified the relevance and correctness of the identified and retrieved literature.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. The human study in Section C.2 involved 20 adult volunteers who gave informed consent before participating, and no personally identifiable information was collected. The human teleoperators shown in our figures and videos consented to being recorded.

REPRODUCIBILITY STATEMENT

We provide detailed information to support the reproducibility of our method and experiments in the supplementary material. Specifically, Section B describes the data representation and network architecture of the motion generator, its pre-training and post-training procedures, the training of the autoregressive motion generator and motion tracker, and the real-world deployment setup. Section C provides dataset statistics and data conversion procedures, the human study setup, definitions of the evaluation metrics, and details on the adaptation and training of baseline methods. Together with the method formulation and experimental setup described in the main paper, these materials specify the key components required to reproduce our results.

REFERENCES

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Videomimic: Visual imitation enables contextual humanoid control. In *Conference on Robot Learning (CoRL)*, 2025. 2

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. In *International Conference on Robotics and Automation (ICRA)*, 2026. 2, 3, 6, 7, S6

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, S3

Maiyue Chen, Kaihui Wang, Bo Zhang, Xihan Ma, Zhiyuan Yang, Yi Ren, Qijun Huang, Zihao Zhu, Yucheng Wang, and Zhizhong Su. Holomotion-1 technical report. *arXiv preprint arXiv:2605.15336*, 2026a. 9

Sirui Chen, Shibo Zhao, Zhen Wu, Jiaman Li, Guanya Shi, and C Karen Liu. Scenebot: Contact-prompted general humanoid whole body tracking with scene-interaction. *arXiv preprint arXiv:2606.27581*, 2026b. 3

Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025. 3

Qiao Feng, Yiming Huang, Yufu Wang, Jiatao Gu, and Lingjie Liu. Physhmr: Learning humanoid control policies from vision for physically plausible human motion reconstruction. In *ACM SIGGRAPH Asia Conference Proceedings*, 2025. 3

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. S8

Sudarshan Harithas, Sangkyung Kwak, Pushkal Katara, Srujan Deolasee, Dvij Kalaria, Srinath Sridhar, Sai Vemprala, Ashish Kapoor, and Jonathan Chung-Kuan Huang. Dreamcontrol-v2: Simpler and scalable autonomous humanoid skills via trainable guided diffusion priors. *arXiv preprint arXiv:2604.00202*, 2026. 3

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. In *Conference on Robot Learning (CoRL)*, 2024a. 3

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *International Conference on Intelligent Robots and Systems (IROS)*, 2024b. 3

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbabu, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi "Jim" Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *Robotics: Science and Systems (RSS)*, 2025. 2

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. S2

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4, S2

Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, S7

Siqiao Huang, Kun-Ying Lee, Dongming Qiao, Guanqi He, Zhenyu Wang, Yitang Li, Shaoting Zhu, and Hang Zhao. Omg: Omni-modal motion generation for generalist humanoid control. *arXiv preprint arXiv:2606.10340*, 2026. 3

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 6, S3

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *ACM SIGGRAPH Asia Conference Proceedings*, 2024a. 6

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. 6

Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. In *Conference on Robot Learning (CoRL)*, 2024c. 3

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023. S1

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *International Conference on Computer Vision (ICCV)*, 2023. 6, S7

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. S2

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and motion estimation from in-the-wild videos. In *International Conference on 3D Vision (3DV)*, 2024. 3

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, 2019. 3

Kyungmin Lee, Sibeen Kim, Minho Park, Hyunseung Kim, Dongyoon Hwang, Hojoon Lee, and Jaegul Choo. Phuma: Physically-grounded humanoid locomotion dataset. In *Conference on Robot Learning (CoRL)*, 2026. 3, S6, S7

Hongjie Li, Heng Yu, Jiaman Li, Hong-Xing Yu, Ehsan Adeli, C Karen Liu, and Jiajun Wu. Anylift: Scaling motion reconstruction from internet videos via 2d diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. 6, S3

Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 6, S3

Jiaman Li, C Karen Liu, and Jiajun Wu. Lifting motion to the 3d world via 2d diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. 6, S3

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *International Conference on Computer Vision (ICCV)*, 2025b. 2, 3, 7, S2, S7

Yuhan Li, Peiyuan Zhi, Yunshen Wang, Tengyu Liu, Sixu Yan, Wenyu Liu, Xinggang Wang, Baoxiong Jia, and Siyuan Huang. Omnitrack: General motion tracking via physics-consistent reference. *arXiv preprint arXiv:2602.23832*, 2026b. S7

Zhe Li, Cheng Chi, Yangyang Wei, Boan Zhu, Tao Huang, Zhenguo Sun, Yibo Peng, Pengwei Wang, Zhongyuan Wang, Fangzhou Liu, et al. Do you have freestyle? expressive humanoid locomotion via audio control. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026c. 3

Zhe Li, Cheng Chi, Boan Zhu, Yangyang Wei, Shuanghao Bai, Yuheng Ji, Yibo Peng, Tao Huang, Pengwei Wang, Zhongyuan Wang, et al. Robomirror: Understand before you imitate for video to humanoid locomotion. In *European Conference on Computer Vision (ECCV)*, 2026d. 3

Zhe Li, Yangyang Wei, Boan Zhu, Yibo Peng, Tao Huang, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Chang Xu, and Shanghang Zhang. From language to locomotion: Retargeting-free humanoid control via motion latent guidance. In *International Conference on Learning Representations (ICLR)*, 2026e. 3

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, 2022. S2

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *Science Robotics*, 11(117):eadx8924, 2026. 3

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 3

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. S3

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. 2, 3, 4, 5, 7, S3, S4

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 6

Massimiliano Pappa, Luca Collorone, Giovanni Ficarra, Indro Spinelli, and Fabio Galasso. Modipo: text-to-motion alignment via ai-feedback-driven direct preference optimization. *Frontiers in Computer Science*, 8:1707808, 2026. 3

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, S7

Zekun Qi, Xuchuan Chen, Dairu Liu, Chenghuai Lin, Yunrui Lian, Sikai Liang, Zhikai Zhang, Yu Guan, Jilong Wang, Wenyao Zhang, et al. Humanoid-gpt: Scaling data and structure for zero-shot motion tracking. *arXiv preprint arXiv:2606.03985*, 2026. 3, 9

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4, S9

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026. 3

Jun Saito, Jiefeng Li, Michael de Ruyter, Miguel Guerrero, Edy Lim, Ehsan Hassani, Roger Blanco Ribera, Hyejin Moon, Magdalena Dadela, Marco Di Lucca, et al. Soma: Unifying parametric human body models. *arXiv preprint arXiv:2603.16858*, 2026. 3, S7

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5, S4, S9

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. S9

Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *ACM SIGGRAPH Asia Conference Proceedings*, 2024. 2, 3, 4, 7, S2, S7

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, S7

Jean Pierre Sleiman, He Li, Alphonsus Adu-Bredu, Robin Deits, Arun Kumar, Kevin Bergamin, Mohak Bhardwaj, Scott Biddlestone, Nicola Burger, Matthew A Estrada, et al. Zest: Zero-shot embodied skill transfer for athletic robot control. *Science Robotics*, 11(117):eaec7695, 2026. 2

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. S2

Bones Studio. Bones-seed: Skeletal everyday embodiment dataset. https://huggingface.co/datasets/bones-studio/seed, 2026. 6

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. S2

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. Sopo: Text-to-motion generation using semi-online preference optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 7, S1

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael Black. Bedlam2. 0: Synthetic humans and cameras in motion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. 6, S3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. S2

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

Yen-Jen Wang, Jiaman Li, Sirui Chen, Takara E Truong, Pei Xu, Pieter Abbeel, Rocky Duan, Koushil Sreenath, Angjoo Kanazawa, Carmelo Sferrazza, et al. Vlk: Learning humanoid loco-manipulation from synthetic interactions in reconstructed scenes. In *Conference on Robot Learning (CoRL)*, 2026a. 3

Yinhuai Wang, Qihan Zhao, Yuen Fui Lau, Runyi Yu, Hok Wai Tsui, Qifeng Chen, Jingbo Wang, Jiangmiao Pang, and Ping Tan. Humanx: Toward agile and generalizable humanoid interaction skills from human videos. In *Conference on Robot Learning (CoRL)*, 2026b. 2

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision (ECCV)*, 2024. 3

Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. S1

Timing Yang, Sicheng He, Hongyi Jing, Jiawei Yang, Zhijian Liu, Chuhang Zou, and Yue Wang. Fast sam 3d body: Accelerating sam 3d body for real-time full-body human mesh recovery. In *European Conference on Computer Vision (ECCV)*, 2026a. S1, S8

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. S1, S8

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

Junpeng Yue, Zepeng Wang, Yuxuan Wang, Weishuai Zeng, Jiangxing Wang, Xinrun Xu, Yu Zhang, Sipeng Zheng, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion models with humanoid control. *arXiv preprint arXiv:2506.12769*, 2025. 3

Yanjie Ze, Zixuan Chen, Joao Pedro Araujo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist: Teleoperated whole-body imitation system. In *Conference on Robot Learning (CoRL)*, 2025. 3

Weishuai Zeng, Kangning Yin, Xiaojie Niu, Shunlin Lu, Weixiang Zhong, Jiahe Chen, Feiyu Jia, Xiao Chen, Zirui Wang, Furui Xu, et al. Scaling behavior foundation model for humanoid robots. *arXiv preprint arXiv:2607.15163*, 2026. 3, 9

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10):12287–12303, 2023. 3

Yangsong Zhang, Anujith Muraleedharan, Rikhat Akizhanov, Abdul Ahad Butt, Gül Varol, Pascal Fua, Fabio Pizzati, and Ivan Laptev. Physmodpo: Physically-plausible humanoid motion with preference optimization. *arXiv preprint arXiv:2603.13228*, 2026a. 3

Youliang Zhang, Ronghui Li, Yachao Zhang, Liang Pan, Jingbo Wang, Yebin Liu, and Xiu Li. A plug-and-play physical motion restoration approach for in-the-wild high-difficulty motions. In *International Conference on Computer Vision (ICCV)*, 2025. 3

Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

Zewei Zhang, Kehan Wen, Michael Xu, Junzhe He, Chenhao Li, Takahiro Miki, Clemens Schwarke, Chong Zhang, Xue Bin Peng, and Marco Hutter. Learning whole-body humanoid locomotion via motion generation and motion tracking. *IEEE Robotics and Automation Letters (RA-L)*, 11(9): 10194–10201, 2026b. 3

Qingrui Zhao, Kaiyue Yang, Xiyu Wang, Shiqi Zhao, Yi Lu, Xinfang Zhang, Qiu Shen, Xiao-Xiao Long, and Xun Cao. Make tracking easy: Neural motion retargeting for humanoid whole-body control. *arXiv preprint arXiv:2603.22201*, 2026a. 3, S6

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and Laszlo A Jeni. Onlinehmr: Video-based online world-grounded human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. 3

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

SUPPLEMENTARY CONTENTS

A Overview **S1**

B Details on ViT-H **S1**

B.1 Data Representation of Motion Generator S1

B.2 Network Architecture of Motion Generator S2

B.3 Pre-Training of Motion Generator S2

B.4 Post-Training of Motion Generator S2

B.5 Adaptation of Autoregressive Motion Generator S4

B.6 Training of Motion Tracker S4

B.7 Real-World Deployment S5

C Details on Experiments **S6**

C.1 Dataset Statistics and Data Conversion S6

C.2 Human Study S7

C.3 Evaluation Metrics S7

C.4 Adaptation of Baselines S7

D Additional Experimental Results **S8**

D.1 Comparisons with Ground-Truth Camera S8

E Limitations and Future Work **S9**

A OVERVIEW

This supplementary material provides further details and additional results for the experiments presented in the main paper. We first describe additional implementation details of **ViT-H** (Section B) and the experimental setup (Section C). We then provide additional experimental results (Section D). We conclude with limitations and directions for future work (Section E). We highly recommend viewing the *supplementary project page* included in the *supplementary materials* for demonstrations across diverse scenarios.

B DETAILS ON VIT-H

In this section, we provide additional details on the data representation (Section B.1), network architecture (Section B.2), pre-training and post-training of our reference motion generator (Sections B.3 and B.4). We further describe the adaptations to the reference motion generator for its autoregressive variant (Section B.5). Finally, we provide details on the training of our motion trackers (Section B.6) and the deployment of our framework in both offline and online settings (Section B.7).

B.1 DATA REPRESENTATION OF MOTION GENERATOR

Video-derived conditions. We represent each video using relative camera orientations, 2D bounding boxes, 2D keypoints, and image features. Specifically, we estimate relative camera orientations with DPVO (Teed et al., 2023), detect human bounding boxes with YOLOv8 (Jocher et al., 2023), extract human keypoints with ViTPose (Xu et al., 2022), and extract image features with Fast SAM 3D Body (Yang et al., 2026a), an accelerated variant of SAM 3D Body (Yang et al., 2026b). For

datasets without paired videos, we simulate camera trajectories, generate 2D bounding boxes, project 3D human motions into the image plane to obtain 2D keypoints, and set the image features to zero.

Global humanoid motion recovery. After predicting the humanoid motion represented by $\mathbf{X} = (\mathbf{v}_{\text{root}}, \mathbf{\Gamma}_{\text{cam}}, \mathbf{\Gamma}_{\text{gv}}, \mathbf{Q}, \boldsymbol{\tau}, \mathbf{p}_{\text{ee}})$ as described in Section 3.1, we discard $\boldsymbol{\tau}$ and $\mathbf{p}_{\text{ee}}$, which are used only for loss computation, and recover the global root transformation following prior human motion reconstruction works that adopt the gravity-view coordinate system (Shen et al., 2024; Li et al., 2025b). We then obtain the joint and vertex positions in world coordinates through forward kinematics using the Unitree G1 kinematic model.

B.2 Network Architecture of Motion Generator

We adopt a Transformer backbone (Vaswani et al., 2017) with rotary positional embeddings (RoPE) (Su et al., 2024) for the video-conditioned reference motion generator π_{gen}. The conditions are independently projected by MLP encoders and then concatenated with the noisy motion before being fed into the backbone. The Transformer outputs are then projected by an output MLP to obtain the denoised motion sequence.

Specifically, the backbone of the reference motion generator consists of 16 Transformer encoder layers with 16 attention heads and a hidden dimension of 1,024, using GELU activations (Hendrycks & Gimpel, 2016). For the model-scaling experiments conducted in Section 4.3, we additionally use a small variant with 12 layers, 8 attention heads, and a hidden dimension of 512, and a medium variant with 14 layers, 12 attention heads, and a hidden dimension of 768.

B.3 Pre-Training of Motion Generator

Training losses. Beyond the diffusion loss described in Equation (2), we introduce the following auxiliary losses to facilitate training: (i) A local joint loss, which applies the Huber loss to root-relative joint positions; (ii) A local vertex loss, which applies the Huber loss to root-relative vertex positions; (iii) a global translation loss, which applies the ℓ_1 loss to the recovered root translation; and (iv) an end-effector stationary loss, which applies binary cross-entropy loss to the stationary states of the hands and feet. For the local vertex loss, we preselect 384 vertices on the Unitree G1 surface using farthest point sampling, with denser sampling around the hands and feet. For the end-effector stationary loss, the ground-truth stationary labels are determined from the velocities of the corresponding joints.

We additionally include the following 2D reprojection losses during training: (i) a 2D joint loss, which applies the Huber loss to reprojected 2D joint positions; (ii) a 2D vertex loss, which applies the Huber loss to reprojected 2D vertex positions; and (iii) a 2D translation loss, which applies the Huber loss to the root translation in camera coordinates following the formulation of CLIFF (Li et al., 2022). For all 2D losses, we use the cameras associated with the paired 3D human motions. We rescale the camera translation by a factor of 0.8 to account for the difference in body scale between humans and the Unitree G1 humanoid.

Training setup. Our video-conditioned reference motion generator is trained using a DDPM scheduler (Ho et al., 2020) with 1,000 diffusion steps. We use the Adam optimizer (Kingma & Ba, 2015) with a fixed learning rate of 2×10^{-4}. For both offline and online generation, we train the model with a fixed sequence length of 120 frames at 30 fps. We train all models on 4 NVIDIA H100 GPUs with a batch size of 64 per GPU, yielding a global batch size of 256. All models are pre-trained for 730k optimization steps.

B.4 Post-Training of Motion Generator

Preference pair construction. For each video-derived condition set $\mathcal{C}$, we construct a candidate pool by sampling $M = 128$ reference motions, $\{\mathbf{X}^{(i)}\}_{i=1}^{M} \sim \pi_{\text{gen}}(\cdot \mid \mathcal{C})$, using DDIM sampling (Song et al., 2021) with 50 denoising steps.

For a humanoid motion sequence $\mathbf{X} = (\mathbf{v}_{\text{root}}, \mathbf{\Gamma}_{\text{cam}}, \mathbf{\Gamma}_{\text{gv}}, \mathbf{Q}, \boldsymbol{\tau}, \mathbf{p}_{\text{ee}})$ that generated by our reference motion generator π_{gen}, we denote the recovered root translation, root orientation, and joint positions

(including the root) in world coordinates as $\mathbf{t} \in \mathbb{R}^{T\times 3}$, $\mathbf{R} \in \mathbb{R}^{T\times 3\times 3}$, and $\mathbf{J} \in \mathbb{R}^{T\times(J+1)\times 3}$, respectively (Section B.1). We further convert the output 6D local joint rotation $\mathbf{Q}$ into 1D joint angle $\mathbf{q} \in \mathbb{R}^{T\times J\times 1}$. For two humanoid motions $\mathbf{X}$ and $\mathbf{X}'$, we compute:

$$\mathcal{E}_{\text{root-pos}}(\mathbf{X},\mathbf{X}') = \frac{1}{T}\sum_t \left\|(\mathbf{t}_t - \mathbf{t}_0) - (\mathbf{t}'_t - \mathbf{t}'_0)\right\|_2^2, \tag{S1}$$

$$\mathcal{E}_{\text{root-ori}}(\mathbf{X},\mathbf{X}') = \frac{1}{T}\sum_t \mathrm{d}^2_{\mathrm{SO(3)}}(\mathbf{R}_t,\mathbf{R}'_t), \tag{S2}$$

$$\mathcal{E}_{\text{pos}}(\mathbf{X},\mathbf{X}') = \frac{1}{T(J+1)}\sum_t\sum_j \left\|\mathbf{R}_t^\top(\mathbf{J}_{t,j} - \mathbf{t}_t) - \mathbf{R}_t'^\top(\mathbf{J}'_{t,j} - \mathbf{t}'_t)\right\|_2^2, \tag{S3}$$

$$\mathcal{E}_{\text{ori}}(\mathbf{X},\mathbf{X}') = \frac{1}{TJ}\sum_t\sum_j (\mathbf{q}_{t,j} - \mathbf{q}'_{t,j})^2, \tag{S4}$$

where $\mathrm{d}_{\mathrm{SO(3)}}$ denotes the geodesic distance on SO(3). We then convert each error into a normalized reward using an exponential kernel. The scale factor in Equation (3) is calibrated from the empirical error distribution as $s_k = \sqrt{\mathrm{Median}(\mathcal{E}_k)}$. We calibrate the scales separately for different data and generation settings to account for differences in their error distributions.

We assign equal weights to all components of the local pose reward $\mathcal{R}_{\text{local}}$ and global motion reward $\mathcal{R}_{\text{global}}$ defined in Equations (4) and (5). For the tracking-success reward $\mathcal{R}_{\text{succ}}$, we determine tracking success following the evaluation termination criterion of SONIC (Luo et al., 2026). Specifically, a rollout is considered failed at frame t if any of the following conditions is met:

$$|z^{\text{ref}}_{\text{root},t} - z^{\text{track}}_{\text{root},t}| > 0.25\,\text{m}, \quad \mathrm{d}_{\mathrm{SO(3)}}(\mathbf{R}^{\text{ref}}_t, \mathbf{R}^{\text{track}}_t) > 1.0\,\text{rad}, \quad \max_{e\in\mathcal{B}_{\text{ee}}} |z^{\text{ref}}_{e,t} - z^{\text{track}}_{e,t}| > 0.25\,\text{m}, \tag{S5}$$

where $\mathcal{B}_{\text{ee}}$ contains the left and right ankle-roll links and the left and right wrist-yaw links. The rollout is considered successful only if no failure is triggered throughout the motion. We apply the same criterion to obtain s and s_{priv} for π_{track} and $\pi^{\text{priv}}_{\text{track}}$, respectively (Equation (6)).

The reward dimensions used for preference-pair construction depend on whether ground-truth humanoid motion is available. In our post-training data, ground-truth humanoid motions are available for BEDLAM (Black et al., 2023), BEDLAM2.0 (Tesch et al., 2026), and Human3.6M (Ionescu et al., 2013), while the video-only datasets (Li et al., 2020; 2025a; 2026a) do not provide ground-truth motions. For videos without ground truth, when evaluating $\mathcal{R}_{\text{global}}(\mathbf{X}, \bar{\mathbf{X}}_{\text{priv}})$ ($r_2^{(i)}$ in Equation (7)), we treat its root and local components as separate reward dimensions during preference-pair filtering.

We impose a margin $\boldsymbol{\delta}$ on the reward advantage when constructing preference pairs. We use different margin settings for the two sources of post-training data. For both data sources, the tracking-success reward uses a zero margin with a non-strict constraint, $r_3^{(i)} - r_3^{(j)} \geq 0$. For the video-only datasets, we set all continuous reward margins to zero to retain as much internet video data as possible. For videos from the pre-training set, which are more abundant, we use the margins to control the amount of preference data. Specifically, we first identify, for each condition, the candidate pair with the largest weighted reward gap under zero continuous margins. We then compute the reward gaps of these selected pairs along each reward dimension and set the corresponding margin δ_k to the p-th percentile of its gap distribution. We tune the shared percentile p such that the number of retained pre-training conditions is comparable to that of the video-only datasets.

We finally select the preference pair from all candidate pairs satisfying the margin constraints by maximizing the weighted reward gap, as described in Equation (9). We assign a weight of 10 to the tracking-success reward and a weight of 1 to all other reward dimensions.

Training setup. For post-training, we set $\beta_{\text{DPO}} = 10$, $\lambda_{\text{DPO}} = 0.05$, and $\lambda_{\text{SFT}} = 2$ for the objective defined in Equations (11) and (12). The SFT objective is applied to the preferred motion $\mathbf{X}^+$. We use the AdamW optimizer (Loshchilov & Hutter, 2019) with a fixed learning rate of 1×10^{-4}, weight decay of 0.01, and gradient norm clipping at 1.0. Unlike pre-training, we do not constrain the motion sequences to a fixed length during post-training. We post-train all models on 4 NVIDIA H100 GPUs with a batch size of 16 per GPU, yielding a global batch size of 64. All models are post-trained for 28k optimization steps.

B.5 Adaptation of Autoregressive Motion Generator

Pre-training setup. For the autoregressive generator $\pi^{\text{ar}}_{\text{gen}}$, we use the same sequence length of $T = 120$ frames and sample the overlap length from $L \in \{0, T-5\}$. When $L = T-5$, the first $T-5$ frames of the noisy motion input are replaced by the corresponding frames from the preceding segment, while $L = 0$ corresponds to generation without preceding-motion conditioning. We use two learnable conditioning tokens to distinguish these two cases, selecting the corresponding token based on whether preceding-motion conditioning is provided. The remaining diffusion scheduler, optimizer, and training configurations follow those of the offline generator (Section B.3).

Post-training setup. For post-training $\pi^{\text{ar}}_{\text{gen}}$, we follow the same DPO objective and training setup (Section B.4). To match the deployment setting (Section B.7), we sample candidate motions using DDIM with 4 denoising steps during post-training. For in-distribution videos from the pre-training datasets with ground-truth motion, we include both samples with and without preceding-motion conditioning. The former use fixed-length sequences of $T = 120$ frames with ground-truth preceding motion, while the latter retain their original sequence lengths without autoregressive conditioning. Internet video samples are similarly used without preceding-motion conditioning and retain their original lengths. We keep the numbers of the two types of pre-training-dataset samples balanced, and their combined duration comparable to that of the Internet video samples.

Autoregressive inference. During online generation, each segment is conditioned on the preceding $T-5$ frames and generates 5 new frames. We maintain a global humanoid motion queue containing all previously generated motions. Before generating $\mathbf{X}^s$, we canonicalize the latest $T-5$ frames in the queue to construct the preceding motion context $\mathbf{X}^{s-1}$. After generation, we align the 5 newly generated frames to the global motion queue using the relative root transformation between the last context frame, i.e., the sixth frame from the end of $\mathbf{X}^s$, and its corresponding frame in the global queue, and then append the aligned frames to the queue. For long-sequence video-to-humanoid teleoperation, we apply a lightweight damping operation to the root transformation of each generated segment to reduce accumulated drift.

B.6 Training of Motion Tracker

Training of the deployable motion tracker. We train π_{track} by fine-tuning SONIC (Luo et al., 2026) with PPO (Schulman et al., 2017). We introduce explicit rewards for tracking the anchor position on the horizontal plane and the anchor yaw orientation, both with a weight of 0.5 and Gaussian kernels with standard deviations of 0.5 m and 0.6 rad, respectively. For PPO optimization, we set the actor learning rate to 5×10^{-6} with an adaptive range of $[5 \times 10^{-6}, 1 \times 10^{-5}]$, and reduce the desired KL divergence from 0.01 to 0.005.

We augment the critic observations with 108 dimensions of privileged global trajectory information, including the current root transformations of both the robot and reference motion and the reference root transformations for the next 10 frames. This increases the critic observation dimension from 1,645 to 1,753 while leaving the actor observation unchanged. To initialize the expanded critic from the released SONIC checkpoint, we copy the weights for the original observations and zero-initialize those associated with the new observations, whose normalization statistics are initialized with zero mean and unit variance. We train π_{track} on 8 NVIDIA H100 GPUs for 20k optimization steps.

Training of the privileged motion tracker. We initialize $\pi^{\text{priv}}_{\text{track}}$ from the trained π_{track} and augment the actor observations with the relative root transformation between the robot and reference motion. Specifically, we append the 3D relative root position and 6D relative root orientation, increasing the actor observation dimension from 930 to 939. We initialize the expanded actor by copying the weights for the original observations and zero-initializing those associated with the new observations. We freeze the SONIC encoders and auxiliary kinematic decoder and optimize only the action decoder and critic. The reward formulation, critic observations, and PPO hyperparameters, including the learning-rate schedule and desired KL divergence, are inherited from π_{track}. We train $\pi^{\text{priv}}_{\text{track}}$ on 8 NVIDIA H100 GPUs for 200k optimization steps.

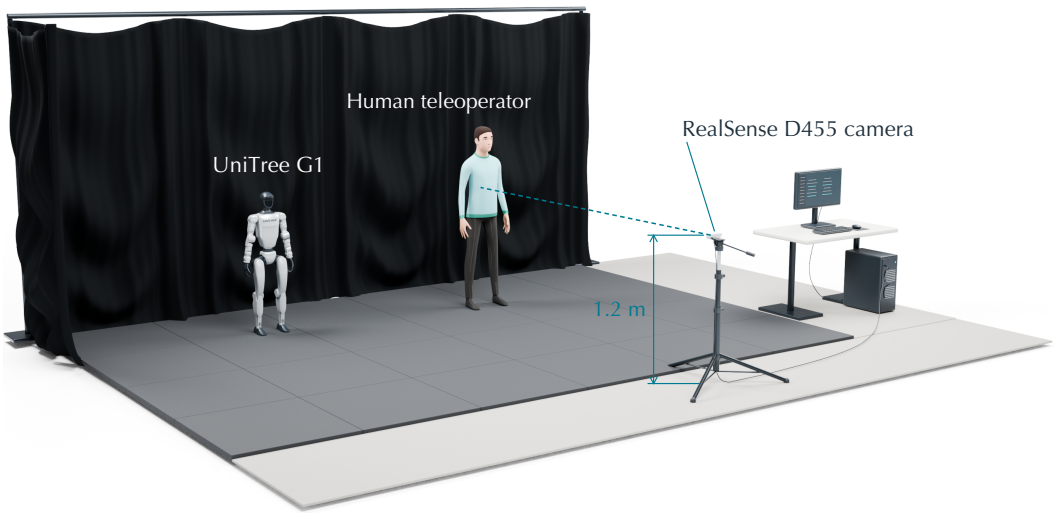


Figure S1: **Video teleoperation setup.** A RealSense D455 camera is fixed 1.2 m above the ground with its optical axis approximately parallel to the ground, capturing RGB video of the human teleoperator. The Unitree G1 humanoid is positioned beside the teleoperator during deployment.

Table S1: **Computational timing breakdown with the NVIDIA RTX 5090 GPU.** We report the average and peak latency of each module in milliseconds. Since peak latency is measured independently for each module, per-module peak latencies are not additive.

Latency	Frame transmission	Video preprocessing	Motion generation	Motion tracking	Total latency
Avg. Latency	70.1	114.4	62.3	6.9	253.7
Peak Latency	130.8	232.1	78.2	12.0	-

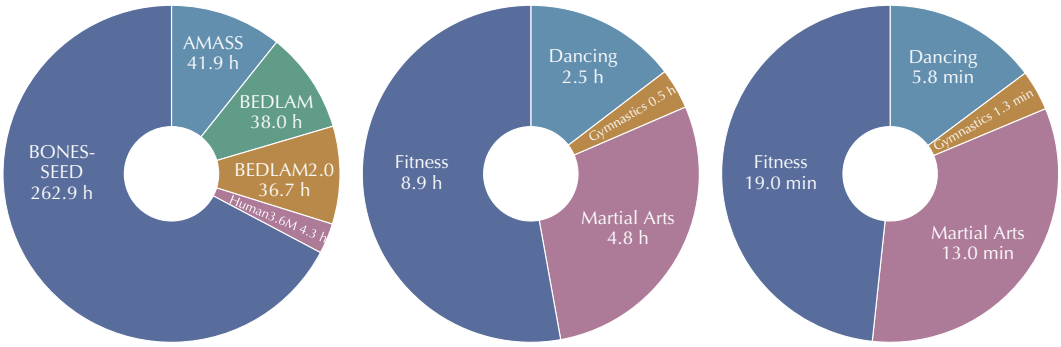
B.7 Real-World Deployment

Deployment setup with offline motion generator. Consistent with our post-training setup (Section B.4), we use DDIM sampling with 50 denoising steps for the offline reference motion generator π_{gen}. We apply no post-processing to the generated motions and directly track them with the deployable motion tracker π_{track} for real-world deployment.

Deployment setup with online motion generator. For real-time video-to-humanoid teleoperation in the online setting, we use DDIM sampling with 4 denoising steps for our autoregressive reference motion generator $\pi^{\text{ar}}_{\text{gen}}$ to reduce inference latency.

We next describe the video teleoperation setup, illustrated in Figure S1. We use an Intel RealSense D455 camera and only use its RGB stream as input, without depth. The camera captures 640×480 images at 30 Hz and is mounted 1.2 m above the ground, with its viewing direction parallel to the ground, to capture the human teleoperator. Each incoming frame is processed online to extract the video-derived conditions (Section B.1). Since the camera remains fixed throughout teleoperation, we do not estimate relative camera orientations.

At system initialization, approximately 10 seconds are required to generate the first motion segment without preceding motion context. Afterwards, the system operates continuously by conditioning on the preceding $T - 5$ frames and generating 5 new frames at a time. Once the preprocessing results for 5 new video frames are available, they are passed to $\pi^{\text{ar}}_{\text{gen}}$ to generate the corresponding reference motion, which is appended to a reference-motion buffer for tracking. The buffered motion is then interpolated to 50 fps before being passed to the deployable tracker π_{track}. Since π_{track} inherits SONIC's 10-frame future reference input, we repeat the current reference frame to avoid additional latency. We provide a computational timing breakdown in Table S1. Our system achieves an average



(a) Pre-training Data Duration by Dataset

(b) Post-training Video-only Data Duration by Category

(c) Test Video-only Data Duration by Category

Figure S2: **Dataset statistics for pre-training, post-training, and testing.** (a) Duration of each dataset used for pre-training. (b) Duration of the video-only post-training data across different motion categories. (c) Duration of the held-out video-only test data across different motion categories.

end-to-end latency below 300 ms. We further reserve a 100 ms latency margin, enabling stable real-time video-to-humanoid teleoperation with sub-400 ms latency.

C Details on Experiments

In this section, we provide additional details on our experimental setup. Throughout our experiments, we use the 29-DoF Unitree G1 as the common humanoid platform for both simulation and real-world deployment. We first present statistics of the pre-training and post-training datasets and describe the data conversion procedures across different representations (Section C.1). We then provide details on the human study setup (Section C.2) and the formulation of our evaluation metrics (Section C.3). We finally detail our adaptations of the baseline methods (Section C.4).

C.1 Dataset Statistics and Data Conversion

Dataset statistics. We summarize the dataset statistics for pre-training, post-training, and testing in Figure S2. For pre-training, we use 383.8 hours of motion data, of which 79 hours are paired with video input, while the remaining data are used without image features. For post-training, our video-only dataset consists of 8,000 videos totaling 16.7 hours and covers diverse motion categories, including dancing, fitness, gymnastics, and martial arts. Our held-out video-only test set contains 282 videos with a total duration of 39.1 minutes.

Adapted GMR retargeter. For both training and testing datasets without paired Unitree G1 motions, we use GMR (Araujo et al., 2026) to retarget the corresponding human motions to the G1 embodiment. We further adapt the original GMR pipeline to improve the robustness and quality of large-scale retargeting.

GMR performs frame-wise inverse kinematics and uses the optimized robot configuration from the previous frame to initialize the next frame. As a result, an inaccurate solution at the beginning of a sequence can propagate throughout the subsequent retargeting. To improve robustness, we run GMR multiple times with different initialization seeds. Each retargeted sequence is filtered and, when necessary, split according to the quality criteria described below, and we retain the result with the largest number of valid frames.

Following prior works on motion curation for humanoid retargeting (Lee et al., 2026; Zhao et al., 2026a), we design a set of rules to filter the source human motions and retargeted robot motions. For human motions, we filter sequences with excessive root jerk or insufficient foot contact. For robot motions, we reject sequences with a minimum root height below 0.25 m, joint velocities exceeding the corresponding hardware limits, severe self-intersection, or excessive foot floating. Specifically, a frame is considered self-intersecting when the penetration depth exceeds 2 cm, and we reject

sequences in which such frames account for more than 5% of the motion; we also require the mean lowest-foot height to remain below 0.10 m.

For the test set, we additionally follow PHUMA (Lee et al., 2026) and apply a base-of-support criterion to the human motion. We construct the support region as the convex hull of the horizontal projections of the left and right ankle and foot joints, and require the average horizontal distances of the pelvis and spine1 from this region to be below 6 cm and 11 cm, respectively.

For sequences containing only localized invalid portions, we split the sequence into shorter segments and retain valid segments with at least 100 frames. Consequently, our adapted GMR retargeter does not necessarily preserve the length of the input sequence, but instead returns the candidate that retains the largest number of valid frames.

Data conversion between human representations. Our humanoid reference motion generator takes human keypoints as input and does not rely on supervision from human body parameters. However, adapting the baseline methods to our training data requires a unified parametric representation of the human body. BONES-SEED represents human motion using SOMA (Saito et al., 2026), whereas all other datasets use SMPL-X (Pavlakos et al., 2019). We therefore convert the human motions in BONES-SEED to the SMPL-X representation.

We first fit the SMPL-X shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ for each actor. We reconstruct the actor-specific SOMA body mesh from its identity, scale, and bone-length parameters. Since the SOMA and SMPL-X models use different rest skeletons and mesh topologies, we re-pose the SOMA mesh using the rest-pose joint configuration of the corresponding SMPL-X model and then map the resulting mesh to the native SMPL-X topology. We then fit $\boldsymbol{\beta}$ to the transferred vertices, using ℓ_2 regularization on the shape parameters and Huber weighting to reduce the influence of large vertex errors. With the fitted $\boldsymbol{\beta}$ fixed, we fit the remaining SMPL-X parameters using the official fitting script provided by SOMA.

C.2 Human Study

We recruit 20 volunteer participants for the human study, all of whom have research backgrounds in computer vision or robotics, including 12 with prior experience in humanoid research. In each trial, participants are shown the original human demonstration video on top and two side-by-side humanoid motion results below. One result is produced by our method and the other by a baseline, with the left-right order randomized. Participants are asked to make a forced-choice judgment on which result better transfers the human motion depicted in the input video to the Unitree G1 humanoid.

C.3 Evaluation Metrics

Metrics for motion quality. To evaluate the quality of the generated humanoid reference motions, we report local MPJPE and PA-MPJPE on root-relative joint positions. We additionally adopt root translation error (RTE) and motion jitter (Jitter) from WHAM (Shin et al., 2024). Self-penetration (Pen.) follows the metric used in OmniTrack (Li et al., 2026b) and is reported in ‰. Details of the 2AFC perceptual study are provided in Section C.2.

Metrics for motion executability. To evaluate the executability of the generated humanoid reference motions, we adopt the standard metrics used in humanoid motion tracking evaluation. The definition of tracking success is provided in Section B.4. For EMDB (Kaufmann et al., 2023) and RICH (Huang et al., 2022), we evaluate tracking success independently on each 250-frame segment and report the average across segments, whereas no fixed segment-length constraint is imposed for the video-only test sets. All remaining metrics are computed on root-relative joint positions. MPJPE and PA-MPJPE are computed between the tracked motion and the retargeted ground-truth motion, while Trk. MPJPE and Trk. PA-MPJPE are computed between the reference motion and its tracked motion.

C.4 Adaptation of Baselines

Adaptations. We adapt the baseline methods, GVHMR (Shen et al., 2024) and GEM (Li et al., 2025b), in terms of their training data, model capacity, and visual encoder. First, we train all baselines on the same datasets used to train our reference motion generator, with the data conversion between

Table S2: **Quantitative comparison of motion quality.** We report results using ground-truth camera parameters provided by the datasets. † denotes adapted baselines trained on our datasets with comparable model sizes. Lower is better for all quantitative metrics. For the 2AFC study, we report the preference rate over Ours (offline).

Method	EMDB					RICH				
	MPJPE	PA-MPJPE	RTE	Jitter	Pen.	MPJPE	PA-MPJPE	RTE	Jitter	Pen.
GVHMR + GMR	73.3	54.4	2.3	17.7	2.3	51.6	44.9	4.6	7.7	1.7
GVHMR† + GMR	71.4	49.1	3.4	15.6	1.6	69.6	51.5	4.4	10.3	2.2
GEM + GMR	62.1	38.3	3.1	18.5	1.2	56.7	49.6	4.7	9.5	2.2
GEM† + GMR	68.8	46.1	3.3	20.5	1.3	70.3	51.2	4.4	20.5	2.2
Ours (online) w/o PT	55.4	36.6	8.3	24.0	0.2	44.5	34.5	6.8	24.4	0.3
Ours (online)	55.4	36.8	6.4	17.5	**0.1**	44.3	36.1	6.3	17.4	0.2
Ours (offline) w/o PT	**52.9**	**34.0**	**1.6**	9.3	0.2	38.7	**28.5**	4.4	**7.2**	0.2
Ours (offline)	53.2	34.2	**1.6**	**8.9**	**0.1**	**38.4**	29.4	**4.3**	**7.2**	**0.1**

Table S3: **Quantitative comparison of motion executability.** We report results using ground-truth camera parameters provided by the datasets. † denotes adapted baselines trained on our datasets with comparable model sizes. Higher is better for Succ.; lower is better for other metrics.

Method	EMDB					RICH				
	Succ.	MPJPE	PA-MPJPE	Trk. MPJPE	Trk. PA-MPJPE	Succ.	MPJPE	PA-MPJPE	Trk. MPJPE	Trk. PA-MPJPE
GT + GMR	91.0	35.2	22.7	35.2	22.7	100.0	30.3	19.6	30.3	19.6
GVHMR + GMR	61.7	96.0	60.0	47.2	29.7	90.3	64.7	47.2	36.4	29.0
GVHMR† + GMR	71.4	93.3	57.2	44.3	29.8	84.7	90.5	58.2	52.5	34.4
GEM + GMR	74.4	84.3	41.5	39.0	23.9	90.3	71.6	51.3	37.8	29.5
GEM† + GMR	75.1	89.2	52.4	41.6	27.6	90.3	88.8	55.1	50.8	33.4
GVHMR	54.9	87.5	46.0	-	-	29.2	83.6	46.4	-	-
GVHMR†	56.4	84.0	45.8	-	-	26.4	82.0	46.5	-	-
GEM	54.9	85.4	45.0	-	-	30.6	86.7	47.7	-	-
GEM†	56.4	83.4	45.4	-	-	23.6	82.0	46.2	-	-
Ours (online) w/o PT	76.2	82.2	41.3	31.7	22.4	87.5	76.8	41.5	35.0	22.2
Ours (online)	87.7	83.1	40.3	**29.4**	**16.4**	98.2	69.8	40.1	27.7	**17.3**
Ours (offline) w/o PT	89.5	73.9	40.1	32.0	18.6	**100.0**	**62.0**	**34.5**	26.2	18.0
Ours (offline)	**90.2**	**71.8**	**39.7**	31.3	18.0	**100.0**	**62.0**	34.6	**25.9**	17.7

different representations described in Section C.1. We set the number of Transformer layers, hidden dimension, and number of attention heads to be the same as those used in our method (Section B.2). Finally, we replace the original HMR 2.0 visual encoder (Goel et al., 2023) with SAM 3D Body (Yang et al., 2026b;a), using the same image features as our method.

Training setup. We train all adapted baselines on 8 NVIDIA H100 GPUs. For each baseline, we follow its original training hyperparameters while keeping the total training budget consistent with our method, using 730k optimization steps.

D Additional Experimental Results

In this section, we provide additional experimental results. We report quantitative comparisons with the baselines and ablation variants using ground-truth cameras (Section D.1). We provide qualitative comparisons against the baselines and ablation variants on the *supplementary project page* included in the *supplementary materials*.

D.1 Comparisons with Ground-Truth Camera

We additionally report quantitative results using ground-truth camera motion for motion quality and motion executability in Tables S2 and S3, respectively. The overall trends remain consistent with those observed under estimated camera motion: our generators outperform the reconstruction-and-retargeting baselines, while post-training further improves motion quality and executability.

E LIMITATIONS AND FUTURE WORK

Limitations. Our experiments demonstrate favorable scaling behavior across the model and data scales considered in this work, for both pre-training and post-training. However, the extent of our scaling study is limited by the available training data and computational resources, and we therefore do not explore substantially larger models or datasets beyond the regime studied here. Additionally, our current motion representation and tracking setup do not model dexterous hand poses. Consequently, the object-interaction behaviors considered in this work are primarily limited to whole-body manipulation of relatively large and simple objects, rather than fine-grained manipulation such as grasping small objects.

Future work. A natural next step is to further scale **ViT-H** with larger models, more diverse pre-training and post-training data, and increased computational resources. Beyond DPO (Rafailov et al., 2023), it would also be interesting to explore alternative post-training algorithms, such as PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024), for directly optimizing motion quality and executability. Moreover, while our experiments show that the improvements from motion-tracking feedback generalize across different trackers, future work could incorporate feedback from multiple trackers during post-training to provide a more robust executability signal. Finally, extending the motion representation and tracking framework to model hand poses would enable finer-grained object interactions, including grasping and manipulation of smaller objects.