# FLOWHMR: PHYSICALLY PLAUSIBLE MOTION CAPTURE FROM VIDEO

**Anonymous authors**
Paper under double-blind review




Figure 1: **Overview of our method.** Given an input video, FlowHMR reconstructs human motion with high fidelity and improved physical plausibility (pink), enabling successful tracking in a physics simulator (blue).

## ABSTRACT

We present FlowHMR, a framework for recovering global 3D human motion from monocular video. Previous methods commonly learn to recover human poses and trajectories through geometric supervision. However, geometric reconstruction objectives do not explicitly account for the dynamics and contact constraints of human motion. Our approach models the video-conditioned distribution of global human motion using Flow Matching. We train the model on a large collection of videos paired with 3D motion annotations and further optimize it through reinforcement learning, jointly rewarding reconstruction fidelity and accurate tracking in simulation. To evaluate motion recovery performance, we introduce Wild-4K, a large and diverse evaluation set of internet videos. Experiments show that FlowHMR outperforms state-of-the-art methods in overall motion quality as judged by human evaluators and in simulation tracking performance. Our code will be made publicly available.

## 1 INTRODUCTION

Recovering global 3D human motion from monocular video is a fundamental task in computer vision and computer graphics (Ye et al., 2023; Shin et al., 2024; Shen et al., 2024). Its applications range from character animation to learning humanoid control from human demonstrations (He et al., 2025; Allshire et al., 2025). The diversity of activities recorded in internet videos makes them a valuable source of motion for these applications, motivating the recovery of faithful and physically plausible human motion in unconstrained environments.

Monocular motion recovery is challenging due to depth ambiguity and occlusion, while image-space motion reflects both human and camera movement. Video-based methods exploit temporal information to recover coherent pose sequences (Kocabas et al., 2020). Global motion recovery further incorporates joint estimation of human and camera motion (Ye et al., 2023), or combines temporal modeling with camera-motion information and contact cues (Shin et al., 2024; Shen et al., 2024).



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.

Learned generative priors have also been incorporated into motion reconstruction from video (Rempe et al., 2021; Li et al., 2025a; Wang et al., 2026b). For learning-based recovery, geometric supervision on image projections, 3D poses, and trajectories establishes correspondence between the predicted motion and the visual observations.

Geometric reconstruction objectives, however, do not explicitly account for the dynamics and contact constraints of human motion. Penalties on foot sliding or ground penetration impose kinematic constraints, but provide only indirect supervision for the underlying dynamics. Physical simulation can complement geometric supervision with feedback on motion execution, and prior work has used physically refined motions to further train motion generators (Gillman et al., 2024; Li et al., 2025c). For video-based recovery, the challenge is to incorporate physical feedback while retaining supervision of the articulation and global trajectory observed in the input video.

We present FlowHMR, a generative framework for recovering global 3D human motion from monocular video (Figure 2). Our approach formulates motion recovery as conditional generation, learning the video-conditioned distribution of global human motion using Flow Matching (Lipman et al., 2023). We represent motion in a unified world coordinate system, jointly encoding body pose, hand articulation, and global trajectories across the sequence. A multimodal diffusion transformer integrates visual features and camera information as conditioning inputs. We train the model on approximately 3,000 hours of synthetic videos paired with 3D motion annotations and further optimize it through physics-based reinforcement learning (RL).

The reinforcement learning objective jointly rewards reconstruction fidelity and accurate tracking in simulation. The fidelity reward measures agreement with the original motion annotations, while the tracking reward evaluates how accurately a fixed humanoid controller follows the generated motion (Luo et al., 2023; 2024). We use MixGRPO (Li et al., 2025b) to update the motion generator with both rewards, incorporating physical feedback while retaining the original annotations as the reconstruction reference.

To evaluate motion recovery performance, we introduce **Wild-4K**, a large and diverse evaluation set of internet videos covering everyday activities, challenging movements, and occlusions. We assess overall motion quality through blinded pairwise human evaluations, complemented by kinematic measurements and simulation tracking. Experiments show that FlowHMR outperforms state-of-the-art methods in overall motion quality as judged by human evaluators and in simulation tracking performance.

Our contributions are threefold:

- A video-conditioned Flow Matching framework that jointly models body pose, hand articulation, and global trajectories in world coordinates for motion recovery from monocular video.
- A physics-based reinforcement learning approach that jointly optimizes reconstruction fidelity and simulation tracking, using the original motion annotations as the fidelity reference.
- Wild-4K, a large and diverse evaluation set of internet videos for assessing motion recovery quality and simulation tracking performance.

## 2 Related Work

**Video-based human motion recovery.** Monocular motion recovery must infer temporally coherent pose and global trajectories despite depth ambiguity, occlusion, and camera motion. HuMoR uses a learned motion prior to constrain reconstruction (Rempe et al., 2021), while GLAMR and SLAHMR combine motion priors with optimization to recover global motion from moving-camera videos (Yuan et al., 2022; Ye et al., 2023). TRACE, WHAM, and GVHMR recover world-space motion through temporal and geometric modeling (Sun et al., 2023; Shin et al., 2024; Shen et al., 2024); PromptHMR extends prompt-conditioned human mesh estimation to video (Wang et al., 2025b). Generative modeling also supports video recovery: GENMO incorporates generative priors into video-based motion estimation (Li et al., 2025a), and DuoMo performs video-conditioned camera-space reconstruction followed by world-space diffusion refinement (Wang et al., 2026b). BEDLAM and BEDLAM2.0 provide paired synthetic videos, body annotations, and camera information for learning reconstruction across varied appearances, motions, and viewpoints (Black et al., 2023;

Tesch et al., 2025). FlowHMR follows this video-to-motion setting, learning a conditional world-space motion distribution from paired video and motion data that supports both reconstruction and subsequent optimization with execution feedback.

**Physics-aware motion recovery and generation.** Physical constraints have been incorporated into reconstruction through physics-based pose optimization in PhysCap (Shimada et al., 2020), differentiable contact and stability terms in IPMAN (Tripathi et al., 2023), and dynamics modeling in D&D and PhysPT (Li et al., 2022; Zhang et al., 2024). SimPoE jointly learns pose refinement and simulated character control (Yuan et al., 2021), while PhysHMR learns a vision-conditioned control policy through expert distillation and reinforcement learning (Feng et al., 2025). Related text-conditioned motion methods integrate physics at different stages: PhysDiff projects intermediate denoised motions in simulation (Yuan et al., 2023); InsActor maps diffusion plans to skills for simulated execution (Ren et al., 2023); CLoSD feeds executed motions back to a diffusion planner (Tevet et al., 2025); and SimDiff trains environment-conditioned adapters on simulated trajectories (Watanabe et al., 2025). FlowHMR uses a fixed PHC+ controller (Luo et al., 2023; 2024) to evaluate sampled motions and jointly rewards their closed-loop execution and framewise geometric agreement with the original annotations. This feedback updates the video-conditioned generator, which produces motions without simulation-based selection or correction at inference.

**Reward-based motion model post-training.** Reward-based optimization has also been applied directly to motion generators. ReinDiffuse fine-tunes motion diffusion with proximal policy optimization (PPO) and rewards for kinematic artifacts (Han et al., 2025), while HY-Motion combines preference optimization and Flow-GRPO with semantic and kinematic rewards (Tencent Hunyuan 3D Digital Human Team, 2025). RLPF uses simulator tracking and alignment rewards to optimize text-conditioned motion generation with group relative policy optimization (GRPO) (Yue et al., 2025); PhysMoDPO uses execution feedback to form preferences over generated motions for diffusion-model optimization (Zhang et al., 2026). For reconstruction, MotionGRPO applies GRPO to full-body recovery from head-mounted-device signals using learned perceptual and joint-level rewards (Yao et al., 2026). We adopt MixGRPO (Li et al., 2025b) as the optimization framework, which builds on GRPO (Shao et al., 2024) and stochastic flow optimization (Liu et al., 2025; Xue et al., 2025). Within this framework, we combine framewise reconstruction against the original motion annotations with simulated tracking feedback for video-conditioned motion recovery, and study their effects on reconstruction fidelity, trackability, and sampling variability.

# 3 Method

## 3.1 Overview

Given a monocular video, FlowHMR recovers global 3D human motion. Figure 2 illustrates our framework, which combines conditional motion generation with physics-based optimization. In Section 3.2, we introduce a Flow Matching model that learns from videos paired with 3D motion annotations. In Section 3.3, we further optimize the model through reinforcement learning, jointly rewarding reconstruction fidelity and accurate tracking in physical simulation.

## 3.2 Generative Video-to-Motion Pretraining

Depth ambiguity and occlusion make monocular motion recovery underdetermined: multiple 3D motions can explain the same visual observations. We therefore formulate recovery as conditional generation and use Flow Matching to learn the distribution of global human motion conditioned on video and camera observations.

### 3.2.1 Conditional Motion Model

Recovering global motion from a moving camera requires a consistent spatial representation across frames. We express the entire sequence in a fixed world coordinate system and encode an $N$-frame motion as $\boldsymbol{x} \in \mathbb{R}^{N \times D}$. Following structured motion representations (Tencent Hunyuan 3D Digital Human Team, 2025; Rempe et al., 2026), each frame includes horizontal root velocity, height, SMPL-H joint positions and continuous 6D rotations (Romero et al., 2017; Zhou et al., 2019), body shape,

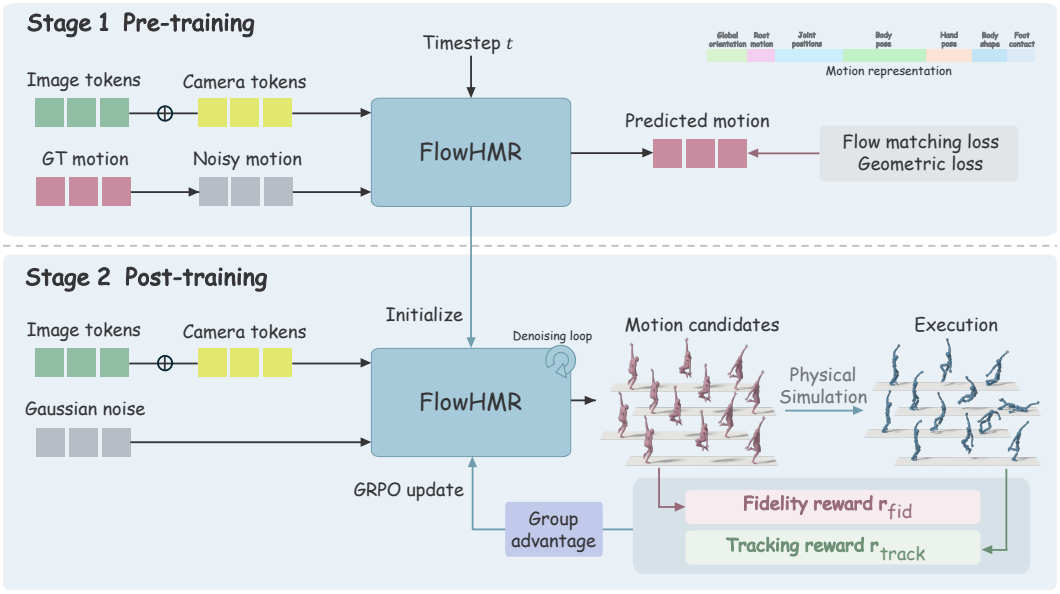



Figure 2: **Pipeline of our method.** FlowHMR consists of two stages for recovering faithful and physically plausible global human motion from monocular video: generative pretraining (Stage 1) and physics-based post-training (Stage 2). In Stage 1, a conditional Flow Matching model learns video-to-motion reconstruction from paired videos and motion annotations with flow and geometric supervision. Starting from the Stage 1 model, Stage 2 further optimizes the generator with GRPO, jointly rewarding fidelity to the original motion annotation and tracking accuracy under a fixed controller in physical simulation.

and foot contacts. This representation allows the model to jointly predict body and hand articulation and the global trajectory.

The model is conditioned on visual observations and camera orientation. We use a frozen pose-aware visual encoder (Yang et al., 2026) to extract features from each input frame. We project the visual features and camera rotations to the same feature dimension and combine them into condition tokens, denoted by $\boldsymbol{c}$.

A Multi-Modal Diffusion Transformer (MMDiT) (Esser et al., 2024; Kong et al., 2024) integrates these conditions with the motion sequence. At flow time $t$, the noisy motion $\boldsymbol{x}_t$ is embedded into motion tokens. Dual-stream blocks use separate projections for motion and condition tokens, with motion tokens attending to visual and camera information. Single-stream blocks then process the concatenated tokens. An embedding of $t$ modulates the transformer blocks, and the output projection predicts clean motion, $\hat{\boldsymbol{x}}_1 = f_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t)$.

In both transformer stages, local temporal attention allows motion tokens to attend to neighboring motion and condition tokens, while condition tokens attend only to neighboring condition tokens. This design combines evidence across frames while keeping the observation features independent of the current noisy motion.

### 3.2.2 Pretraining with Flow Matching

We train the conditional generator with Flow Matching (Lipman et al., 2023), learning to transform Gaussian noise into motion using paired video and 3D motion annotations. For each training pair, let $\boldsymbol{x}_1$ denote the annotated motion and $\boldsymbol{c}$ the video and camera conditions. We interpolate between Gaussian noise $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ to obtain a noisy motion $\boldsymbol{x}_t$. Using the model's clean-motion prediction (Li & He, 2025), we define the predicted and target flow fields as

$$\begin{aligned} \boldsymbol{x}_t &= (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \qquad \boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \\ \widehat{\boldsymbol{v}}_\theta &= \frac{f_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t) - \boldsymbol{x}_t}{d(t)}, \qquad \boldsymbol{v}^* = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_t}{d(t)}, \end{aligned} \tag{1}$$

where $t \in [0, 1]$ and $d(t) = \max(1 - t, \varepsilon_t)$ stabilizes the denominator. Componentwise Smooth-L1 losses between these fields define $\mathcal{L}_{\text{FM}}$. We additionally supervise joint rotations and positions decoded from the clean prediction through forward kinematics:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{rot}}\mathcal{L}_{\text{fk_rot}} + \lambda_{\text{pos}}\mathcal{L}_{\text{fk_pos}}. \tag{2}$$

The flow loss learns to transform noise into motion conditioned on the observations, while the forward-kinematics losses directly constrain the resulting joint geometry.

Motion generation follows the learned flow from Gaussian noise using Euler integration. We decode the resulting representation into SMPL-H motion, recovering the global trajectory from horizontal root velocities and pelvis positions and the articulated body from joint rotations and shape.

### 3.3 Physics-based Post-training

The geometric objectives above supervise motion reconstruction without explicitly accounting for dynamics and contact constraints. We further optimize the generator using feedback from physical simulation. A fixed controller tracks each generated motion, and its tracking accuracy provides a reward for reinforcement learning. We combine this feedback with a reconstruction fidelity reward to preserve correspondence to the observed motion.

#### 3.3.1 Candidate Sampling and Evaluation

Given a video condition $\boldsymbol{c}$ and its ground-truth motion annotation $\boldsymbol{x}^{\text{gt}}$, we sample a group of $G$ complete motions $\{\hat{\boldsymbol{x}}_i\}_{i=1}^{G}$ using the stochastic flow sampler described in Section 3.3.2. Each candidate is evaluated for fidelity to the original annotation and tracking accuracy in simulation.

**Fidelity Reward.** Optimizing tracking alone can favor motions that are easy to execute but depart from the action in the input video. We therefore retain the original paired motion annotation as the fidelity reference throughout post-training. For any sampled motion $\hat{\boldsymbol{x}}$, we decode it and the annotation into SMPL-H joint positions, denoted by $\hat{\boldsymbol{q}}_{n,j}$ and $\boldsymbol{q}^{\text{gt}}_{n,j}$ for frame $n$ and joint $j$. The fidelity reward is

$$r_{\text{fid}} = \exp\left[-\frac{1}{\tau L J}\sum_{n=1}^{L}\sum_{j=1}^{J}\left\|\hat{\boldsymbol{q}}_{n,j} - \boldsymbol{q}^{\text{gt}}_{n,j}\right\|_2\right], \tag{3}$$

where $L$ is the valid sequence length, $J$ is the total number of body and hand joints, and $\tau$ controls the reward scale. We compare corresponding frames in the same sequence coordinate frame without additional pose alignment, retaining sensitivity to both articulation and global trajectory. This supplies a reconstruction preference for body and hand motion during physical optimization, retaining the original paired annotation as the reconstruction target.

**Tracking Reward.** To assess whether a candidate can be followed under dynamics and contact constraints, we convert it to the rigid-body representation of a fixed PHC+ controller (Luo et al., 2023; 2024), which attempts to track it in simulation. The generated motion itself serves as the tracking reference. Let $\boldsymbol{b}^{\text{ref}}_{n,m}$ denote its reference position for rigid body $m$ and $\boldsymbol{b}^{\text{exec}}_{n,m}$ the corresponding position in the simulated rollout. The tracking reward is

$$r_{\text{track}} = \exp\left[-\frac{1}{\tau H}\sum_{n=1}^{H}\min\left(\frac{1}{J_{\text{sim}}}\sum_{m=1}^{J_{\text{sim}}}\left\|\boldsymbol{b}^{\text{exec}}_{n,m} - \boldsymbol{b}^{\text{ref}}_{n,m}\right\|_2, d_{\max}\right)\right], \tag{4}$$

where $H$ is the scored rollout length and $J_{\text{sim}}$ is the number of controller rigid bodies. The per-frame cap $d_{\max}$ limits the contribution of extreme deviations before temporal averaging. The resulting sequence score provides graded feedback on how closely the executed trajectory follows the generated reference. This evaluation adds feedback from simulated dynamics to the annotation-based fidelity signal, encouraging accurate execution under the fixed controller and environment.

#### 3.3.2 Reward-Guided Policy Optimization

We initialize the generator from pretrained weights and adopt MixGRPO (Li et al., 2025b), denoted GRPO in our comparisons, to learn from complete-motion rewards without differentiating through simulation.

At each iteration, the current generator samples $G$ complete motions per video. Candidates share initial Gaussian noise and receive independent perturbations within the stochastic sampling window; outside it, sampling follows deterministic Euler updates.

We retain only groups containing both successful and failed tracking outcomes. Within each group, the two rewards are standardized separately, weighted, summed, and clipped to obtain candidate advantages. These guide policy updates only at the sampled stochastic transitions. Only the generator is updated; the visual encoder and controller remain fixed. Appendix C.2 details the sampler and policy objective.

# 4 Experiments

We evaluate motion capture from internet videos through blinded human comparisons, supported by simulation tracking and geometric measurements.

## 4.1 Experimental Setup

**Datasets.** We pretrain FlowHMR on approximately 3,000 hours of synthetic video rendered from 650 hours of motion capture and animation assets; data construction and motion preprocessing are detailed in Appendices A and B. We introduce Wild-4K, an evaluation set of 4,182 internet video clips spanning everyday motion, self-occlusion, and difficult actions. These clips have no ground-truth motion annotations. We use the full set for tracking and geometric evaluation. For human evaluation, we sample comparison pairs without replacement from a pool constructed using 500 Wild-4K videos; the evaluated video sets differ across baselines. RICH (Huang et al., 2022) provides annotated reconstruction diagnostics for the ablations.

**Human evaluation.** Our primary evaluation uses blinded pairwise comparisons of direct predictions for the same video, with randomized A/B order. Annotators judge overall motion quality through naturalness, video fidelity, global trajectory, and ground-contact artifacts. Finger poses use a shared default and are excluded from assessment. G/S/B denote wins, ties, and losses for our method, aggregated over valid annotations; Appendix D.2 details the protocol.

**Evaluation metrics.** To complement the overall human judgment, we report *Success Rate* under a fixed PHC+ controller, with *Tracking Score* additionally reported in the post-training ablations. Following PhysDiff (Yuan et al., 2023), we measure *Penetration*, *Floating*, and *Skating* on generated motions before simulation. These quantify displacement below the ground, displacement above the ground, and horizontal foot motion during contact, respectively; we report the median per-sequence error for each metric.

**Compared methods.** We compare with GVHMR, PromptHMR, GENMO, and DuoMo (Shen et al., 2024; Wang et al., 2025b; Li et al., 2025a; Wang et al., 2026b). *Ours w/o RL* and *Ours* denote FlowHMR before and after post-training.

**Inference settings.** FlowHMR generates one motion per video without simulation-based selection or motion correction. Wild-4K tracking evaluation uses up to 12 seconds per clip; complete sampling and evaluation settings are provided in Appendix D.

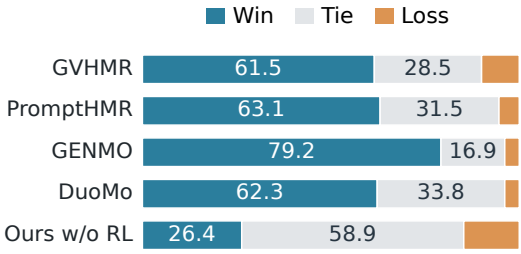
## 4.2 Results on Internet Videos

### 4.2.1 Blinded Human Evaluation

In Table 1 (left), FlowHMR wins 61.5–79.2% of valid annotations against the four external methods, with losses of 3.8–10.0%.

The comparison with Ours w/o RL evaluates the perceptual effect of post-training within the same model. Among 129 valid annotations, the post-trained model receives 34 wins (26.4%), 76 ties (58.9%), and 19 losses (14.7%). Post-training is preferred more often than the pretrained model, while the majority of comparisons are judged equal.

Table 1: **Comparison on Wild-4K. Left:** our method outperforms all baselines in blinded human evaluation, with more wins than losses in every comparison. Bars show win/tie/loss percentages over valid judgments; evaluated videos differ across comparisons. **Right:** simulation Success Rate (SR, %) under PHC+ and physical errors on 4,182 clips. Our method achieves the highest SR, the lowest penetration (Pen.) and floating (Float) errors, and the second-lowest skating (Skate) error. Physical errors are median per-sequence values (mm) measured before simulation. Throughout, bold/underlined values mark best/second-best results; ties share formatting.




| Method | SR ↑ | Pen. ↓ | Float ↓ | Skate ↓ |
|---|---|---|---|---|
| GVHMR | 62.82 | 1.02 | 0.81 | **1.59** |
| PromptHMR | 60.33 | 1.10 | 2.71 | 3.64 |
| GENMO | 58.80 | 1.52 | 1.20 | 2.55 |
| DuoMo | 56.71 | 1.76 | 2.08 | 2.56 |
| Ours w/o RL | <u>78.79</u> | <u>0.60</u> | <u>0.53</u> | 2.48 |
| Ours | **82.47** | **0.06** | **0.20** | <u>2.17</u> |

#### 4.2.2 Tracking and Physical Plausibility

On Wild-4K (Table 1, right), Ours w/o RL reaches a Success Rate of 78.79%, compared with the highest external baseline rate of 62.82%. Post-training raises this rate to 82.47%, a gain of 3.68 percentage points.

To determine whether the tracking gains are accompanied by fewer geometric defects, we examine the generated motions before simulation. Median Penetration decreases from 0.60 to 0.06 mm, Floating from 0.53 to 0.20 mm, and Skating from 2.48 to 2.17 mm after post-training. GVHMR still has lower Skating at 1.59 mm, showing that the best tracking result does not imply the lowest value for every geometric measurement.

### 4.3 Ablation Studies

#### 4.3.1 Pretraining Design

We study how training data, generative modeling, and model size affect reconstruction on the internal Studio set. Studio contains 300 video sequences spanning diverse motion categories with optical motion capture ground truth, supporting quantitative comparisons of body, hand, and trajectory reconstruction. Table 2 reports mean results across its motion categories.

Table 2: **Ablations of the pretrained model.**

| Factor | Setting | WA-MPJPE (mm) ↓ | MPJPE (mm) ↓ | RTE (mm) ↓ | $\text{MPJRE}_{\text{hand}}$ (°) ↓ | Jitter ($\text{mm/s}^2$) ↓ | Foot-sliding (mm) ↓ |
|---|---|---|---|---|---|---|---|
| Data scale | 200 h | <u>149.44</u> | <u>90.92</u> | <u>4.91</u> | <u>25.17</u> | <u>16.34</u> | <u>3.53</u> |
| | 3,000 h | **67.82** | **43.49** | **1.86** | **16.08** | **11.75** | **2.55** |
| Prediction | Regression | <u>73.28</u> | **43.44** | <u>2.39</u> | **14.55** | <u>18.01</u> | <u>3.94</u> |
| | Flow Matching | **67.82** | <u>43.49</u> | **1.86** | <u>16.08</u> | **11.75** | **2.55** |
| Model size | 0.2B | 67.82 | 43.49 | 1.86 | <u>16.08</u> | 11.75 | 2.55 |
| | 0.5B (selected) | <u>63.39</u> | **42.31** | **1.81** | 16.48 | <u>11.03</u> | **2.27** |
| | 1B | **62.84** | <u>42.37</u> | <u>1.83</u> | **15.96** | **10.93** | <u>2.28</u> |

**Training data scale.** We evaluate the effect of pretraining data scale by comparing models trained on 200 and 3,000 hours of pretraining data. The 3,000-hour model improves all six reported metrics. WA-MPJPE decreases from 149.44 to 67.82 mm and hand-rotation error from 25.17 to 16.08 degrees.

**Generative formulation.** We compare Flow Matching with feed-forward regression to assess the effect of generative modeling on reconstruction. Flow Matching reduces WA-MPJPE from 73.28 to

Table 3: **Ablation of post-training.** We compare different post-training methods and reinforcement-learning reward settings. Our GRPO with joint fidelity (Recon.) and tracking (Sim.) rewards achieves higher simulation success than both SFT variants while keeping reconstruction errors close to the pretrained model. Optimizing tracking alone yields high success at the cost of severe reconstruction errors. Adding a kinematic proxy reward (KinProxy: penetration, floating, and skating) gives similar success but worse reconstruction than our two-reward design. * marks reward hacking, excluded from ranking.

| Method | Target / reward | RICH | | Wild-4K (PHC+) | |
|---|---|---|---|---|---|
| | | WA-MPJPE (mm) ↓ | PA-MPJPE (mm) ↓ | Success (%) ↑ | Score ↑ |
| Pretrained | — | <u>55.70</u> | **30.78** | 78.79 | 0.692 |
| SFT | Filtered originals | **54.80** | <u>30.88</u> | 78.38 | 0.690 |
| | Projected targets | 70.80 | 42.13 | 80.25 | 0.706 |
| GRPO | Recon. only | 57.29 | 31.62 | 73.79 | 0.658 |
| | Sim. only* | 255.67 | 139.20 | 99.62* | 0.942* |
| | Recon. + KinProxy | 62.15 | 37.43 | 77.98 | 0.671 |
| | Recon. + Sim. + KinProxy | 60.66 | 34.38 | **82.57** | <u>0.744</u> |
| **Ours** | Recon. + Sim. | 56.68 | 31.40 | <u>82.47</u> | **0.746** |

67.82 mm and also improves RTE, Jitter, and Foot-sliding. MPJPE is similar (43.49 versus 43.44 mm), while hand-rotation error is higher (16.08 versus 14.55 degrees).

**Model size.** Increasing model size from 0.2B to 1B reduces WA-MPJPE from 67.82 to 62.84 mm and Jitter from 11.75 to 10.93. The gains are not monotonic across all metrics: the 0.5B model has the lowest MPJPE, RTE, and Foot-sliding among the three sizes.

### 4.3.2 Post-training Design

Table 3 compares supervision targets and reward combinations using the pretrained model as a shared reference. Reconstruction and tracking are reported together to assess whether tracking gains preserve the recovered action. A synchronized qualitative comparison is provided in Figure C.1.

**Post-training strategies.** We test whether supervised fine-tuning (SFT) with filtered original targets or simulated trajectories obtains tracking gains comparable to joint-reward GRPO. SFT with filtered originals retains the original motion targets and reduces RICH WA-MPJPE to 54.80 mm, but does not improve Wild-4K tracking over the pretrained model. Projected-target SFT uses simulated body poses and root trajectories as supervision targets. It increases Success Rate to 80.25%, but raises WA-MPJPE to 70.80 mm and PA-MPJPE to 42.13 mm. Joint-reward GRPO reaches 82.47% Success Rate with smaller reconstruction costs, retaining the original annotation as its fidelity reference.

**Reward components.** The fidelity and tracking rewards serve different objectives, so we first test what happens when either is used alone. Fidelity-only GRPO does not improve reconstruction or tracking over the pretrained model. Tracking-only GRPO reaches 99.62% Success Rate, but WA-MPJPE and PA-MPJPE increase to 255.67 and 139.20 mm. This reward-hacking behavior shows why high tracking scores must be assessed together with reconstruction fidelity.

The kinematic proxy reward (KinProxy) combines the Penetration, Floating, and Skating metrics defined above to favor motions with fewer geometric artifacts. Replacing the tracking reward with a kinematic proxy gives a Success Rate of 77.98%, below the pretrained model's 78.79%.

Adding the kinematic proxy to the joint fidelity and tracking rewards yields a similar Success Rate (82.57% versus 82.47%), but higher reconstruction errors and a slightly lower Tracking Score. We use the two-reward configuration for its better reconstruction at comparable tracking performance.

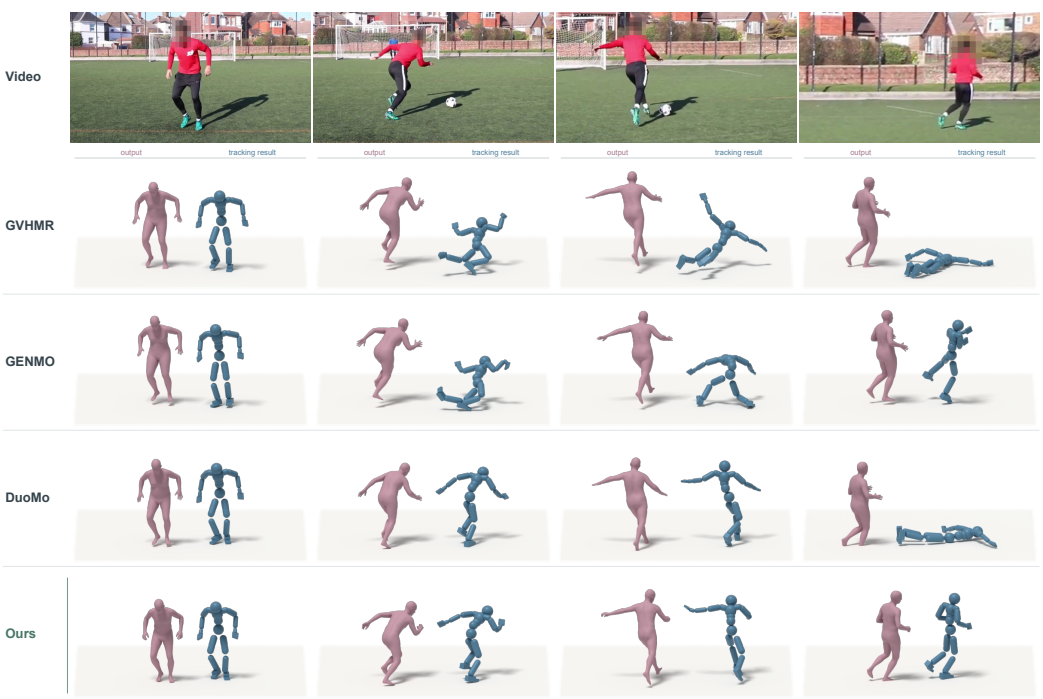



Figure 3: **Qualitative comparison of reconstruction and simulation tracking.** The top row shows the input video, followed by GVHMR (Shen et al., 2024), GENMO (Li et al., 2025a), DuoMo (Wang et al., 2026b), and FlowHMR. Each method panel pairs the reconstructed output (pink, left) with its tracking result in simulation (blue, right). Our method enables successful simulation tracking of this sequence, whereas all compared baselines fail.

### 4.4 Qualitative Comparisons

In Figure 3, GVHMR and DuoMo fall during PHC+ tracking, while GENMO deviates substantially from its generated reference. FlowHMR remains upright with torso orientation and arm poses closer to the reference. Visually plausible reconstruction alone does not guarantee stable physical execution. Additional qualitative examples and comparisons appear in Figures C.1, D.1 and D.2.

## 5 Limitations and Future Work

Physics-based simulation provides feedback that improves the physical plausibility and tracking of recovered motion. However, physical execution is currently evaluated in a simplified environment containing only a ground plane. Simulation-based post-training also requires repeated rollouts for sampled motions, making training costly as the data scale, sequence length, or number of samples grows.

Internet videos often contain interactions with objects and support surfaces. Future work could reconstruct these entities and model their contacts in simulation, while developing more efficient simulation and rollout strategies to scale physical feedback to more videos and longer sequences.

## 6 Conclusion

We presented FlowHMR, a generative framework for recovering global 3D human motion from monocular video. It combines Flow Matching pretraining with physics-based reinforcement learning to jointly optimize reconstruction fidelity and simulation tracking. We also introduced Wild-4K and found improved overall motion quality and tracking performance over the evaluated baselines.

### AI USE STATEMENT

Generative AI assisted with manuscript organization and editing, organization of experimental summaries, numerical consistency checks, and LaTeX preparation.

### REPRODUCIBILITY STATEMENT

The main paper and appendices describe the method, data processing, training settings, and evaluation protocols. We will publicly release the code.

## REFERENCES

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of the Conference on Robot Learning*, 2025. URL https://arxiv.org/abs/2505.03729.

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

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 3 2024. URL http://arxiv.org/abs/2403.03206.

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 *SIGGRAPH Asia 2025 Conference Papers*. Association for Computing Machinery, 2025. doi: 10.1145/3757377.3763951. URL https://fengq1a0.github.io/projects/physhmr/index.html.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. URL https://arxiv.org/abs/2107.08430.

Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. Self-correcting self-consuming loops for generative model training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15646–15677. PMLR, 2024. URL https://proceedings.mlr.press/v235/gillman24a.html.

Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. ReinDiffuse: Crafting physically plausible motions with reinforced diffusion model. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 2218–2227, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Han_ReinDiffuse_Crafting_Physically_Plausible_Motions_with_Reinforced_Diffusion_Model_WACV_2025_paper.html.

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 K. Hodgins, Linxi Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.066. URL https://www.roboticsproceedings.org/rss21/p066.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13274–13285, 2022.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5253–5263, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Kocabas_VIBE_Video_Inference_for_Human_Body_Pose_and_Shape_Estimation_CVPR_2020_paper.html.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. Technical report, Tencent, 12 2024. URL http://arxiv.org/abs/2412.03603.

Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&D: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2209.08790.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. MixGRPO: Unlocking flow-based GRPO efficiency with mixed ODE-SDE. *arXiv preprint arXiv:2507.21802*, 2025b. URL https://arxiv.org/abs/2507.21802.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Zhuo Li, Mingshuang Luo, Ruibing Hou, Xin Zhao, Hao Liu, Hong Chang, Zimo Liu, and Chen Li. Morph: A motion-free physics optimization framework for human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14580–14589, 2025c. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_Morph_A_Motion-free_Physics_Optimization_Framework_for_Human_Motion_Generation_ICCV_2025_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025. URL https://arxiv.org/abs/2505.05470.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2305.06456.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.04582v2.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. doi: 10.1126/scirobotics.aed4592. URL https://www.science.org/doi/abs/10.1126/scirobotics.aed4592.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Int. Conf. Comput. Vis.*, pp. 5442–5451, 2019.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. URL https://arxiv.org/abs/2108.10470.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

QuickMagic. QuickMagic. https://www.quickmagic.ai/, n.d. Commercial motion capture platform. Accessed September 26, 2026.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11488–11499, October 2021.

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:2603.15546*, 2026.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, Xiao Ma, Liang Pan, and Ziwei Liu. InsActor: Instruction-driven physics-based characters. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/bc943cd038a5531d5433b1431c822c01-Abstract-Conference.html.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36, 11 2017. ISSN 15577368. doi: 10.1145/3130800.3130883.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

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 *SIGGRAPH Asia Conference Proceedings*, 2024.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM Transactions on Graphics*, 39(6), 2020. URL https://vcai.mpi-inf.mpg.de/projects/PhysCap/.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8856–8866, 2023.

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

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J Black. Bedlam2.0: Synthetic humans and cameras in motion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Haim Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZISppZSTv.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4713–4725, 2023. URL https://arxiv.org/abs/2303.18246.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://vggt-omega.github.io/.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36M: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8428–8437, 2025a.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1148–1159, 2025b.

Yufu Wang, Evonne Ng, Soyong Shin, Rawal Khirodkar, Yuan Dong, Zhaoen Su, Jinhyung Park, Kris Kitani, Alexander Richard, Fabian Prada, and Michael Zollhöfer. DuoMo: Dual motion diffusion for world-space human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42777–42788, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_DuoMo_Dual_Motion_Diffusion_for_World-Space_Human_Reconstruction_CVPR_2026_paper.html.

Akihisa Watanabe, Jiawei Ren, Siyao Li, Yichen Peng, Erwin Wu, and Edgar Simo-Serra. SimDiff: Simulator-constrained diffusion model for physically plausible motion generation. *arXiv preprint arXiv:2509.20927*, 2025. URL https://arxiv.org/abs/2509.20927.

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*, volume 35, pp. 38571–38584, 2022.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. URL https://arxiv.org/abs/2505.07818.

Xitong Yang, Devansh Kukreja, Don Pinkus, et al. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

Nanjie Yao, Junlong Ren, Wenhao Shen, and Hao Wang. MotionGRPO: Overcoming low intra-group diversity in GRPO-based egocentric motion recovery. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.05680.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21222–21232, 2023.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7159–7169, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yuan_SimPoE_Simulated_Character_Control_for_3D_Human_Pose_Estimation_CVPR_2021_paper.html.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11038–11049, 2022.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.02500.

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. URL https://arxiv.org/abs/2506.12769v1.

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*, 2026. URL https://arxiv.org/abs/2603.13228.

Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. PhysPT: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2305–2317, 2024. URL https://arxiv.org/abs/2404.04430.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.*, pp. 5745–5753, 6 2019. URL http://arxiv.org/abs/1812.07035.

# A Data Processing

## A.1 Synthetic Data for Pretraining

We construct paired videos and 3D motion annotations by combining motion assets, procedural cameras, and randomized scene rendering. Approximately 650 hours of source motion yield more than 3,000 hours of rendered video. Motion capture collections and professional animation assets cover locomotion, dance, yoga, sports, stylized actions, hand articulation, and global trajectories. Sequences are converted to neutral SMPL-H (Romero et al., 2017): SMPL (Loper et al., 2015) and SMPL-X (Pavlakos et al., 2019) motions undergo mesh fitting, and other skeletons are retargeted. Quality control removes duplicates, screens joint-limit violations and interpenetration, prunes static segments, and flags foot sliding; these checks do not certify physical executability. Rendering the same motion under varied observations complements synthetic supervision resources such as AMASS, BEDLAM, and BEDLAM2.0 (Mahmood et al., 2019; Black et al., 2023; Tesch et al., 2025).

**Camera trajectories.** We sample static, subject-following, orbital, arc, and dolly-style cameras, favoring static and tracking views. Focal lengths span 18–200 mm with greater probability in the 32–60 mm range; camera heights cover low, eye-level, high, and aerial views, while azimuth sampling favors frontal and side views. Subject bounding boxes guide camera distance, with framing that includes both full-body views and partial truncation at image boundaries. Dynamic cameras follow smoothed skeleton-derived positions and orientations, with optional height variation and low-frequency position/rotation perturbations to model handheld motion.

**Rendering and appearance.** Blender's GPU-accelerated Cycles renderer independently samples camera motion, character textures, ground materials, and HDR environment maps to vary appearance and illumination. Explicit ground geometry receives shadows and uses scanned material channels. Geometric primitives with opaque, transparent, metallic, or glass materials create partial occlusions at varied image locations. Optional motion blur varies the virtual shutter. Videos are rendered at 1080p and 30 fps with paired motion and camera annotations; Figure A.1 shows examples. These visual variations do not enforce dynamically valid contact in the underlying motion.



(a) Foreground occlusion (b) Challenging pose (c) Lower-body truncation (d) Lateral truncation

Figure A.1: Synthetic pretraining examples with (a) foreground occlusion, (b) a challenging pose, and (c,d) partial visibility caused by image-boundary truncation.

## A.2 Data Filtering for Post-training

We select original video–motion pairs using PHC+ tracking feedback, retaining motions with a tracking score above 0.6 and without early termination. The resulting post-training pool contains approximately 571k pairs. Filtering retains the original motion annotations as reconstruction targets, rather than substituting simulated trajectories. The GRPO loader reuses supervised target construction, disables augmentation and condition dropout, and retains the first 180 frames of longer inputs while padding shorter sequences. Projected-target SFT separately substitutes exported controller-tracked body motion while retaining the paired video conditions; its conversion and export procedure is specified in Appendix C.

### A.3 WILD-4K EVALUATION SET ACQUISITION

We construct Wild-4K from Koala-36M (Wang et al., 2025a), using a vision-language model (VLM), ViTPose (Xu et al., 2022), and YOLOX (Ge et al., 2021) to screen for clips featuring **a single adult with full-body framing**. **We exclude non-ground-based settings, including underwater, aerial, and suspended scenes**, yielding 4,182 clips. Seed2.0 Lite classifies each retained video into one of three evaluation groups: *normal*, covering everyday motions such as walking and speaking; *difficult*, covering actions such as gymnastics and exercise; and *self-occluded*, covering back-facing views, camera-view occlusion, and foreground occlusion. Full-body framing does not require every body part to remain visible throughout the clip. We then estimate camera poses for each video using VGGT-Ω (Wang et al., 2026a). The same clip set and VLM-assigned groups are used across methods; Appendix D describes the evaluation protocol.

## B GENERATIVE PRETRAINING

We supplement Section 3.2 with representation, architecture, and training details.

### B.1 MOTION REPRESENTATION AND DECODING

Motion is expressed in a sequence reference derived from the initial camera orientation, with initial horizontal translation removed. Each frame comprises

$$\boldsymbol{x}_i = [\, \boldsymbol{u}_i^{xz},\ h_i,\ \mathrm{vec}(\boldsymbol{P}_i),\ \mathrm{vec}(\boldsymbol{Q}_i),\ \boldsymbol{\beta}_i,\ \boldsymbol{b}_i \,], \tag{5}$$

where, following Kimodo (Rempe et al., 2026), $\boldsymbol{u}_i^{xz}$ and $h_i$ encode horizontal root velocity and absolute root height separately. $\boldsymbol{P}_i$ and $\boldsymbol{Q}_i$ are joint positions and continuous 6D rotations (Zhou et al., 2019), and $\boldsymbol{\beta}_i$ and $\boldsymbol{b}_i$ encode shape and foot contact. Features are standardized using training-set statistics. Root velocity is the frame-rate-scaled difference of a smoothed pelvis trajectory. Joint positions subtract this trajectory horizontally while retaining absolute height and the sequence orientation. Root rotation uses the sequence reference; other rotations are relative to their parents.

Decoding integrates horizontal velocities from a zero origin and adds the predicted pelvis offset. Vertical position comes from the pelvis-position channel; the separate height channel supplies auxiliary supervision. Shape coefficients are averaged over time, and SMPL-H translation subtracts the resulting rest-pose pelvis offset from the recovered pelvis position. Predicted and annotated motions are decoded using their respective shapes, with joint positions in meters. Contact labels, derived from low foot velocity and proximity to the ground, provide auxiliary supervision.

### B.2 MMDIT ARCHITECTURE AND CONDITIONING

The MMDiT backbone (Kong et al., 2024) uses six dual-stream blocks followed by twelve single-stream blocks, with hidden width 1,024, 16 attention heads, and MLP expansion ratio 4. Dual-stream blocks use separate motion and condition projections; single-stream blocks process their concatenation. Frozen SAM 3D Body features (Yang et al., 2026) and relative-camera-rotation features are separately projected and summed, then contextualized by a two-layer condition refiner. Temporal rotary embeddings encode frame order, and flow-time embeddings modulate the blocks.

Local attention uses a broad motion context and a narrower observation window. In both block types, motion tokens attend to neighboring motion and condition tokens, whereas condition tokens attend only to neighboring conditions. This asymmetric mask keeps observation features independent of the noisy motion tokens.

### B.3 OBJECTIVES, OPTIMIZATION, AND SAMPLING

**Training objectives.** We use the interpolation and clean-motion parameterization in Equation 1, with $t = \mathrm{sigmoid}(z)$ and $z \sim \mathcal{N}(\mu_t, \sigma_t^2)$. Let $P_k$ select representation component $k$. Field supervision is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\sum_{k \in \mathcal{C}} \alpha_k \,\mathrm{SmoothL1}\left(P_k(\widehat{\boldsymbol{v}}_\theta), P_k(\boldsymbol{v}^*)\right)\right], \tag{6}$$

where $\alpha_k$ balance trajectory, articulation, shape, and contact components. For geometric supervision, clean predictions and targets are decoded separately using their own rotations, shape, and translation. Forward kinematics produces global joint rotations $\boldsymbol{G}$ and positions $\boldsymbol{J}$:

$$\begin{aligned} \mathcal{L}_{\text{fk_rot}} &= \text{SmoothL1}\left(\text{rot6d}(\widehat{\boldsymbol{G}}), \text{rot6d}(\boldsymbol{G}^*)\right), \\ \mathcal{L}_{\text{fk_pos}} &= \text{SmoothL1}(\widehat{\boldsymbol{J}}, \boldsymbol{J}^*). \end{aligned} \tag{7}$$

The geometric losses use no pelvis or Procrustes alignment. All losses average over feature coordinates and valid frames, excluding padding, and combine as in Equation 2.

**Training and inference.** We pretrain for 500,000 steps with a global batch size of 1,024, using AdamW with cosine learning-rate decay from $10^{-4}$ to $10^{-5}$. Gradients are clipped to a maximum $\ell_2$ norm of 10. Visual-feature sequences are randomly dropped during training while camera features are retained. Inference uses 20 uniform Euler steps from Gaussian noise without classifier-free guidance, followed by SMPL-H decoding. The stochastic sampling used for post-training is specified in Appendix C.

### B.4 Pretraining Evaluation

We evaluate the pretrained video-to-motion model through reconstruction metrics and human ratings. Throughout this section, "Ours" denotes the pretrained model.

**Studio evaluation.** Studio is an internal evaluation set of 300 video sequences spanning diverse motion categories, with ground-truth motion obtained through optical motion capture. MPJPE and $\text{MPJRE}_{\text{hand}}$ denote mean per-joint position error and mean per-joint rotation error for the hands, respectively.

#### B.4.1 Sampling Variability

Each of 32 generation seeds is evaluated on the complete Studio set. Table B.1 reports the seed with the lowest WA-MPJPE, the mean across seeds, and the seed with the highest WA-MPJPE. The gap between the best and mean results characterizes variation in reconstruction quality within the pretrained sampling distribution, suggesting scope for preference optimization to favor better outputs.

Table B.1: **Sampling variability of the pretrained model.** Each of 32 seeds is evaluated on all 300 Studio sequences. Best/Worst report all metrics from the seed with the lowest/highest dataset-level WA-MPJPE. Mean averages each metric across seeds.

| Seed | WA-MPJPE↓ (mm) | MPJPE↓ (mm) | RTE↓ (mm) | $\text{MPJRE}_{\text{hand}}$ ↓ (°) | Jitter↓ ($\text{mm/s}^2$) | Foot-sliding↓ (mm) |
|---|---|---|---|---|---|---|
| Best | **60.62** | **41.09** | **1.72** | **14.42** | **10.82** | **2.22** |
| Mean | <u>63.18</u> | <u>43.51</u> | <u>1.79</u> | <u>15.80</u> | <u>10.97</u> | <u>2.24</u> |
| Worst | 66.19 | 46.39 | 1.88 | 17.06 | 11.10 | 2.27 |

#### B.4.2 Blind Human Evaluation

Human evaluation on Wild-300 covers six categories: combat and martial arts, dance performance, field sports, fitness exercise, daily activities, and locomotion. Videos were manually selected for visible, trackable subjects, with unequal numbers of clips across categories. Each method's recovered motion was rendered alongside or overlaid on its input video. Five or more evaluators with experience assessing motion quality rated outputs independently, with method identities hidden and presentation order randomized. Each of six dimensions used a 5-point Likert scale. Figure B.1 reports mean ratings weighted by the number of clips in each category. Table B.2 summarizes the scoring criteria.

**Human evaluation scoring rubric.** Evaluators compare each reconstruction with its input video using the six criteria in Table B.2. Shared score anchors are **5**: faithful, natural motion with negligible

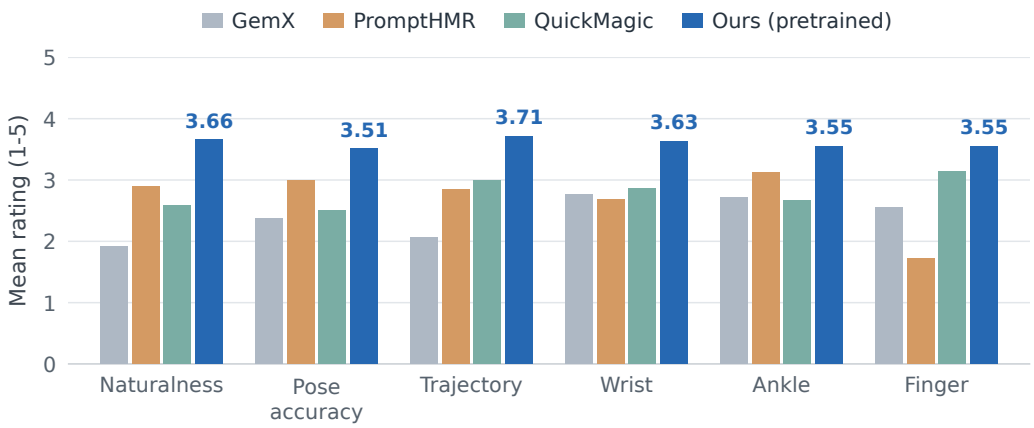



Figure B.1: **Human evaluation results on Wild-300.** Comparisons include QuickMagic (n.d.). Bars show mean ratings weighted by the number of clips in each category. Ratings use a five-point scale; higher is better.

defects; **3**: recognizable motion with limited deviations; and **1**: substantially incorrect, missing, or implausible motion. Scores 2 and 4 interpolate between adjacent anchors.

Table B.2: **Human evaluation criteria on Wild-300.** Each dimension is scored independently for the pretrained models.

| **Dimension** | **Assessment criteria** |
|---|---|
| Core pose accuracy | Torso and limb orientations, depth, spine curvature, and plausible balance should match the video without body interpenetration. Wrists, ankles, and fingers are assessed separately. |
| Overall naturalness and fluidity | Assess temporal continuity, weight and inertia, speed variation, anticipation, and follow-through. Penalize jitter, abrupt stops, foot sliding, floating, and ground penetration. |
| Global trajectory and orientation | Compare the overall path, facing direction, travel distance, and vertical displacement. Reduced movement is a deviation; incorrect direction or absent locomotion is a severe error. |
| Ankle angle | Compare foot orientation and ankle flexion and rotation. Distinguish moderate angular deviations from incorrect orientation or anatomically implausible twisting. |
| Wrist angle | Compare palm direction, wrist rotation, and flexion or extension. Distinguish moderate angular deviations from reversed orientation or implausible joint bending. |
| Finger precision | Compare gestures, finger articulation, and opening or closing dynamics. Penalize inaccurate bend angles, missing articulation, interpenetration, and implausible joint configurations. |

# C Post-training Details and Analysis

## C.1 PHC+ Motion Tracking

We use a frozen PHC+ controller (Luo et al., 2023; 2024) to track generated motions in Isaac Gym (Makoviychuk et al., 2021) during post-training. Each motion is converted to the controller's reference format at 30 fps with upright-start alignment and $y$-up to $z$-up conversion. The controller starts from the reference state and tracks the motion in simulation, without recovery, fall initialization, or motion cycling.

The rollout supplies the executed rigid-body positions used in Equation 4. For an $N$-frame reference, the scored horizon is $H = \max(1, N-1)$; the tracking-error cap and reward scale are listed in Table C.1. During post-training, invalid reference conversions or unavailable simulation results receive zero tracking reward, while the fidelity reward remains defined.

### C.2 MixGRPO Implementation

Our post-training follows MixGRPO (Li et al., 2025b), including its mixed ODE–SDE sampler, windowed policy updates, and separate group normalization of multiple rewards. We detail their implementation for the motion generator below.

**Stochastic transitions.** Sampling advances from noise to motion on the grid $t_k = k/S$, $k = 0, \dots, S-1$, with $h = 1/S$ and $\sigma_k = 1 - t_k$. We define

$$\boldsymbol{v}_{\theta,k} = \frac{f_\theta(\boldsymbol{x}^k, \boldsymbol{c}, t_k) - \boldsymbol{x}^k}{\max(\sigma_k, h)}, \qquad a_k = \eta\sqrt{\frac{\sigma_k}{\max(t_k, h)}}, \qquad \nu_k = a_k\sqrt{h}. \tag{8}$$

Here $\eta$ controls exploration. Within the stochastic window,

$$\begin{aligned} \boldsymbol{\mu}_{\theta,k} &= \left(1 - \frac{a_k^2 h}{2\sigma_k}\right)\boldsymbol{x}^k + h\left(1 + \frac{a_k^2 t_k}{2\sigma_k}\right)\boldsymbol{v}_{\theta,k}, \\ \boldsymbol{x}^{k+1} &\sim \mathcal{N}(\boldsymbol{\mu}_{\theta,k}, \nu_k^2 \boldsymbol{I}). \end{aligned} \tag{9}$$

Outside the window, sampling follows $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + h\boldsymbol{v}_{\theta,k}$. Candidates for the same video share initial noise and receive independent noise within the window.

**Transition scores.** Let $p_\theta(\boldsymbol{x}^{k+1} \mid \boldsymbol{x}^k, \boldsymbol{c})$ denote the Gaussian transition density in Equation 9. For $d = DT$ coordinates, including padded frames, the dimension-normalized Gaussian score is

$$\bar{\ell}_{\theta,i,k} = -\frac{\|\boldsymbol{x}_i^{k+1} - \boldsymbol{\mu}_{\theta,k}\|_2^2}{2d\nu_k^2} - \log\nu_k - \tfrac{1}{2}\log(2\pi) = d^{-1}\log p_\theta(\boldsymbol{x}_i^{k+1} \mid \boldsymbol{x}_i^k, \boldsymbol{c}). \tag{10}$$

Here $D$ is the motion dimension and $T$ is the padded sequence length. These scores define the dimension-normalized likelihood ratio used for the policy updates in Section 3.3.2.

**Window schedule.** The window $\mathcal{W}_s = \{s, \dots, \min(s+W, M) - 1\}$ spans at most $W$ of the first $M$ transitions. Its start advances by one periodically and cycles through $s = 0, \dots, M-1$; windows are truncated at the boundary. We denote the current window by $\mathcal{W}$.

**Group selection.** Before computing advantages, we exclude any video group whose candidates all succeed or all fail at tracking. Only groups with both tracking outcomes contribute to the RL update. All candidates in a retained group, including failed rollouts, participate in advantage computation. This group-level selection is separate from the per-component variance safeguard below.

**Advantage normalization.** Each reward component is standardized within its video group using the sample standard deviation and stabilizer $\varepsilon_s$. A component receives zero normalized advantage if its group has fewer than two candidates or its standard deviation is nonfinite or below $\delta$. Following MixGRPO's multi-reward aggregation, we combine the two components as

$$A_i = \mathcal{C}_A(w_{\text{fid}}\mathcal{N}_g(r_{\text{fid},i}) + w_{\text{track}}\mathcal{N}_g(r_{\text{track},i}))\,, \tag{11}$$

where $i$ indexes candidates for the same video, $\mathcal{N}_g$ denotes the group standardization above, $\mathcal{C}_A$ clips to $[-A_{\max}, A_{\max}]$, and $w_{\text{fid}}$ and $w_{\text{track}}$ weight the two components.

**Policy objective.** We collect candidate motions and their stochastic transitions using the sampling parameters $\theta_{\text{old}}$. For the retained video groups, the transition scores above give a dimension-normalized likelihood ratio for the clipped policy objective (Schulman et al., 2017; Li et al., 2025b):

$$\begin{aligned} \rho_{i,k}(\theta) &= \left[\frac{p_\theta(\boldsymbol{x}_i^{k+1} \mid \boldsymbol{x}_i^k, \boldsymbol{c})}{p_{\theta_{\text{old}}}(\boldsymbol{x}_i^{k+1} \mid \boldsymbol{x}_i^k, \boldsymbol{c})}\right]^{1/d}, \\ \mathcal{J}(\theta) &= \mathbb{E}\left[\frac{1}{G|\mathcal{W}|}\sum_{i=1}^{G}\sum_{k\in\mathcal{W}}\min(\rho_{i,k}A_i, \text{clip}(\rho_{i,k}, 1-\epsilon, 1+\epsilon)A_i)\right]. \end{aligned} \tag{12}$$

Here $\epsilon$ is the likelihood-ratio clipping threshold, and the expectation is over retained video groups. Maximizing $\mathcal{J}$ applies each complete motion's advantage $A_i$ to its stochastic transitions in $\mathcal{W}$.

We partition the stochastic transitions from retained groups into at most $K$ disjoint minibatches, recompute their scores on the stored states, and take one optimizer step per minibatch. The objective averages over transitions, candidates, and retained videos; each selected transition is used once per iteration. Collected states, old scores, and advantages remain fixed throughout these updates.

### C.3 Training Settings

Post-training initializes the generator from the pretrained weights. Only the generator is optimized; the visual encoder and controller remain fixed. We use AdamW with cosine learning-rate decay and gradient clipping, without additional old-score or pretrained-field regularization. Table C.1 reports the settings used for joint-reward GRPO.

Table C.1: **GRPO hyperparameters.** Sampling, reward, and optimization settings used for joint-reward post-training. The video batch size $B$ is measured before group selection.

| Setting | Value |
|---|---|
| Videos per device $B$ / candidates per video $G$ | 4 / 8 |
| Sampling steps $S$ / maximum sequence length | 20 / 180 frames |
| Noise strength $\eta$ | 0.7 |
| Window length $W$ / eligible transitions $M$ | 4 / 8 |
| Window shift interval | 100 outer iterations |
| Updates per outer iteration (maximum $K$) | 4 |
| Initial / minimum learning rate | $2.5 \times 10^{-6}$ / $2.5 \times 10^{-7}$ |
| Gradient norm limit / surrogate clip $\epsilon$ | 1 / $10^{-3}$ |
| Group standard-deviation floor $\delta$ / stabilizer $\varepsilon_s$ | $10^{-4}$ / $10^{-8}$ |
| Advantage bound $A_{\max}$ | 5 |
| Reward scale $\tau$ / tracking-error cap $d_{\max}$ | 0.15 m / 0.5 m |
| Fidelity:tracking weights | 1 : 1 |

**Projected-target SFT.** We select nonterminated references with tracking score above 0.6 and at least 30 scored steps. Simulated body rotations and root translation replace the corresponding annotations; original hand rotations, shape, video features, and camera conditions are retained. We restore the original vertical offset after coordinate conversion and repeat the last exported frame when needed to match the source length.

### C.4 Sampling Variability

We evaluate fixed pretrained and final GRPO checkpoints with eight seeds on a 300-video subset of Wild-4K: 153 normal, 99 self-occluded, and 48 difficult motions. Table C.2 reports variation among generated motions and execution quality under PHC+, following the metric definitions in Appendix D.

Table C.2: **Sampling before and after post-training.** Ours w/o RL and Ours are each evaluated with eight seeds on a 300-video subset of Wild-4K. (a) Pairwise seed MPJPE averages the 28 pairs per video, then takes the median across videos. Tracking MPJPE averages the eight seed-level errors between execution and generated references over all motions. (b) Mean $\pm$ standard deviation of the eight seed-level PHC+ results.

**(a) Motion variation and tracking**

| Model | Pairwise seed MPJPE (mm) | Tracking MPJPE (cm) $\downarrow$ |
|---|---|---|
| Ours w/o RL | 57.25 | <u>8.35</u> |
| Ours | 22.86 | **6.96** |

**(b) Execution across seeds**

| Model | Success Rate (%) $\uparrow$ | Tracking Score $\uparrow$ |
|---|---|---|
| Ours w/o RL | <u>76.33</u> $\pm$ 1.02 | <u>0.6799</u> $\pm$ 0.0055 |
| Ours | **79.71** $\pm$ 0.42 | **0.7344** $\pm$ 0.0019 |

Table C.2 shows better mean tracking performance and lower sensitivity to sampling seeds after GRPO. Reduced variation across samples is distinct from the movement amplitude within a motion sequence.

### C.5 Qualitative Post-training Comparison

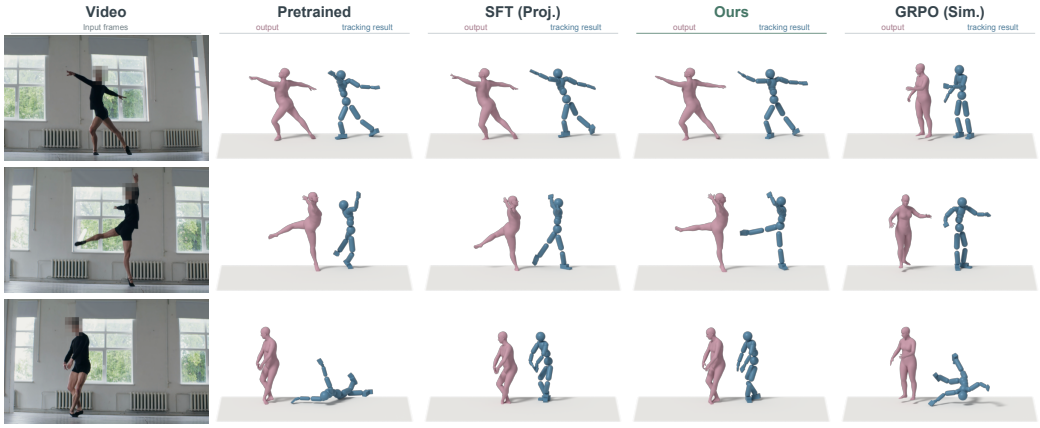



Figure C.1: **Qualitative comparison of post-training strategies.** Columns show the input video, pretrained FlowHMR, SFT with projected targets, joint-reward GRPO (Ours), and tracking-only GRPO (Sim.). Rows show three synchronized time points. Each method panel pairs the reconstructed output (pink, left) with its tracking result in simulation (blue, right).

## D Additional Evaluation

### D.1 Evaluation Protocol

**Evaluation datasets.** **Wild-4K** is a large and diverse evaluation set curated from publicly available video data. The full set is used for tracking and geometric evaluation. A 300-video subset is used for sampling and post-processing analyses. G/S/B comparisons are sampled without replacement from a pool constructed using 500 Wild-4K videos, as detailed in Section D.2. **RICH** (Huang et al., 2022) is a public academic benchmark used to evaluate reconstruction accuracy and PHC+ execution. **Studio** contains 300 sequences spanning diverse motion categories with optical motion capture ground truth, and is used primarily for iterative model development. **Wild-300** is a manually curated video set used primarily for human evaluation and qualitative comparisons.

**Reconstruction and inference.** We evaluate reconstruction on RICH using world-aligned, world-coordinate, and Procrustes-aligned mean per-joint position error (WA-/W-/PA-MPJPE), root translation error (RTE, %), jitter ($\mathrm{m/s^3}$), and foot sliding. RTE and jitter are evaluated over the whole sequence. FlowHMR uses one generated motion per video, with seed 0, 20 sampling steps, and no classifier-free guidance or additional motion correction. Sampling variability is examined in Section C.4.

**Motion evaluation.** PHC+ evaluation runs in Isaac Gym and uses all 4,182 Wild-4K clips described in Appendix A.3, with at most 360 frames per clip at 30 Hz. The fixed controller supplies both training feedback and the reported execution evaluation. Its tracking score is $r_{\mathrm{track}}(\hat{\boldsymbol{x}})$, and success requires

$$\mathrm{pass}(\hat{\boldsymbol{x}}) = \mathbf{1}[r_{\mathrm{track}}(\hat{\boldsymbol{x}}) > \gamma_{\mathrm{pass}} \ \wedge \ \neg\,\mathrm{terminated}(\hat{\boldsymbol{x}}) \ \wedge \ \mathrm{survival}(\hat{\boldsymbol{x}}) = 1]\,, \tag{13}$$

with $\gamma_{\mathrm{pass}} = 0.6$, where survival is the fraction of the scored horizon completed. For Wild-4K, Success Rate uses all 4,182 evaluated clips as its denominator. Following PhysDiff (Yuan et al., 2023), Penetration and Floating measure the lowest mesh vertex's distance below and above the ground, with a 5 mm tolerance; Skating measures horizontal foot displacement between consecutive frames with contact. Each error is averaged within a sequence, in millimetres. The Penetration, Floating, and Skating columns in Table 1 report the median of each metric over all 4,182 generated motions before simulation. Skating on Wild-4K and foot sliding on RICH retain their respective evaluation protocols. PHC+ Tracking MPJPE compares simulated execution with the generated reference and includes all evaluated motions, whether successful or not.

### D.2 Human Evaluation on Internet Videos

**Protocol.** We select 500 videos from Wild-4K and pair FlowHMR with each of the five comparison methods on every video, forming a pool of 2,500 comparison pairs. We sample pairs without replacement for annotation; each sampled pair is evaluated once by a single annotator. The resulting evaluated video sets differ across comparison methods. Each judgment presents the original video and two direct motion predictions for that video, before simulation. The predictions are labeled A and B, with method identities hidden and presentation order randomized; evaluators select A, B, or a tie. The overall judgment considers four criteria: motion naturalness, motion accuracy relative to the input video, global trajectory, and foot sliding, penetration, or floating. Finger poses are set to the same default for all methods and are excluded from the assessment. A tie records no overall preference; the four criteria are not scored separately.

Table D.1: **G/S/B human evaluation on Wild-4K.** G/S/B denote post-trained FlowHMR's wins/ties/losses against each method, including its pretrained counterpart (Ours w/o RL). We report 649 valid judgments and 13 technical skips; $n = G + S + B$ excludes skips. Win and win-or-tie rates are $G/n$ and $(G + S)/n$, respectively.

| Compared method | G | S | B | $n$ | Win (%) | Win-or-tie (%) | Skipped |
|---|---|---|---|---|---|---|---|
| GVHMR | 80 | 37 | 13 | 130 | 61.5 | 90.0 | 2 |
| PromptHMR | 82 | 41 | 7 | 130 | 63.1 | 94.6 | 1 |
| GENMO | 103 | 22 | 5 | 130 | 79.2 | 96.2 | 1 |
| DuoMo | 81 | 44 | 5 | 130 | 62.3 | 96.2 | 3 |
| Ours w/o RL | 34 | 76 | 19 | 129 | 26.4 | 85.3 | 6 |

**Aggregation.** Each pair contributes one vote; percentages therefore use valid judgments rather than unique videos. The Wild-300 evaluation in Figure B.1 separately assesses pretrained models using a five-point rating scale.

### D.3 Reconstruction and Physical Execution on RICH

Despite the limitations of existing benchmarks, we also report quantitative reconstruction and physical execution results on RICH (Huang et al., 2022) to facilitate comparison with prior methods.

Table D.2: **Quantitative reconstruction and physical-execution results on RICH.** WA/W/PA denote WA-/W-/PA-MPJPE; these and foot sliding (FS) use mm. RTE is in %; Jitter in $\mathrm{m/s^3}$. PHC+ Success (%) and tracking MPJPE (mm) use all 185 sequences, including failures; the latter compares execution with the generated reference motion. Ours achieves the highest PHC+ Success Rate and the lowest tracking MPJPE, although its reconstruction errors are not uniformly the lowest.

| Method | Reconstruction | | | | Motion quality | | PHC+ execution | |
|---|---|---|---|---|---|---|---|---|
| | WA ↓ | W ↓ | PA ↓ | RTE ↓ | Jitter ↓ | FS ↓ | Success ↑ | MPJPE ↓ |
| GVHMR | 82.83 | 133.08 | 40.05 | 2.49 | 20.69 | 3.64 | 28.11 | 154.3 |
| PromptHMR | 72.25 | 115.55 | 41.92 | <u>1.71</u> | 19.56 | 4.27 | 35.14 | 150.5 |
| GENMO | 88.12 | 140.51 | 42.20 | 2.49 | 16.44 | 4.02 | 35.68 | 148.2 |
| DuoMo | **52.86** | **78.28** | 37.43 | **1.33** | 7.53 | <u>3.29</u> | 24.86 | 179.1 |
| Ours w/o RL | <u>55.70</u> | <u>88.37</u> | **30.78** | 1.92 | <u>7.07</u> | 3.63 | <u>48.65</u> | <u>133.5</u> |
| Ours | 56.68 | 88.43 | <u>31.40</u> | 1.82 | **6.71** | **3.14** | **51.35** | **118.8** |

### D.4 EXECUTION UNDER SONIC

We retarget Wild-4K reconstructions to the Unitree G1 humanoid using GMR (Araujo et al., 2025) and evaluate SONIC (Luo et al., 2026) tracking in Isaac Lab (Mittal et al., 2025); Figure D.3 shows an execution example. A clip is successful when its reference sequence finishes without a non-timeout termination. We use the 615-clip success-union subset formed from 3,247 clips evaluated for all five methods, and report success rates on this subset in Table D.3, rather than on the full Wild-4K set. FlowHMR reaches a 77.72% success rate on this subset.

Table D.3: **SONIC success rate on a 615-clip Wild-4K subset.** Bold and underlining mark the best and second-best values.

| Method | Success rate (%) ↑ |
|---|---|
| PromptHMR | 44.07 |
| GENMO | 48.78 |
| DuoMo | 51.06 |
| GVHMR | <u>54.15</u> |
| FlowHMR (ours) | **77.72** |

### D.5 More Qualitative Results

**We strongly encourage readers to watch the supplementary videos for a more complete assessment of motion quality and physical execution.**

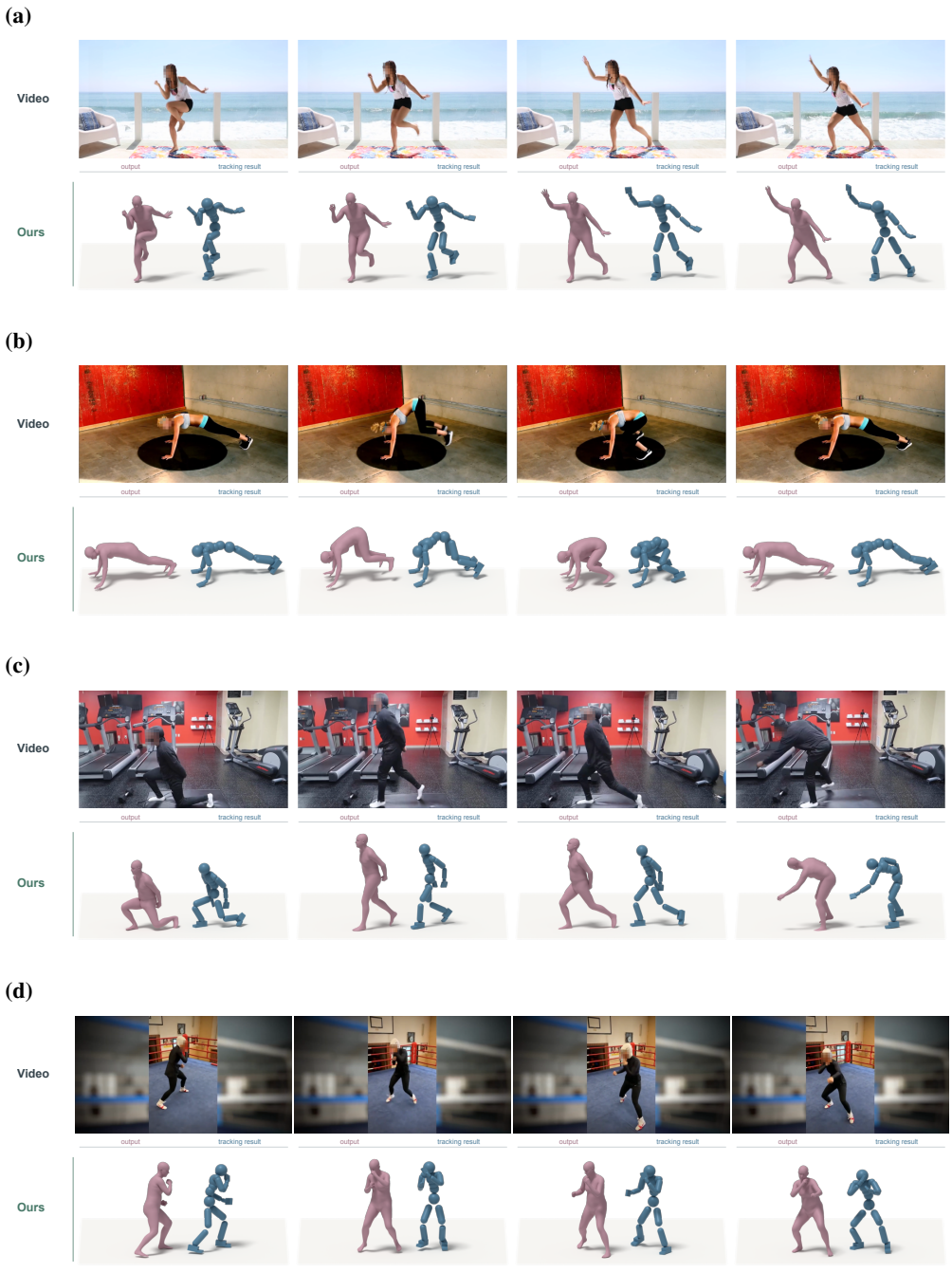



Figure D.1: **Additional qualitative examples.** The upper row shows the input video, and the lower row pairs the motion generated by Ours (pink, left) with its tracking result in simulation (blue, right).

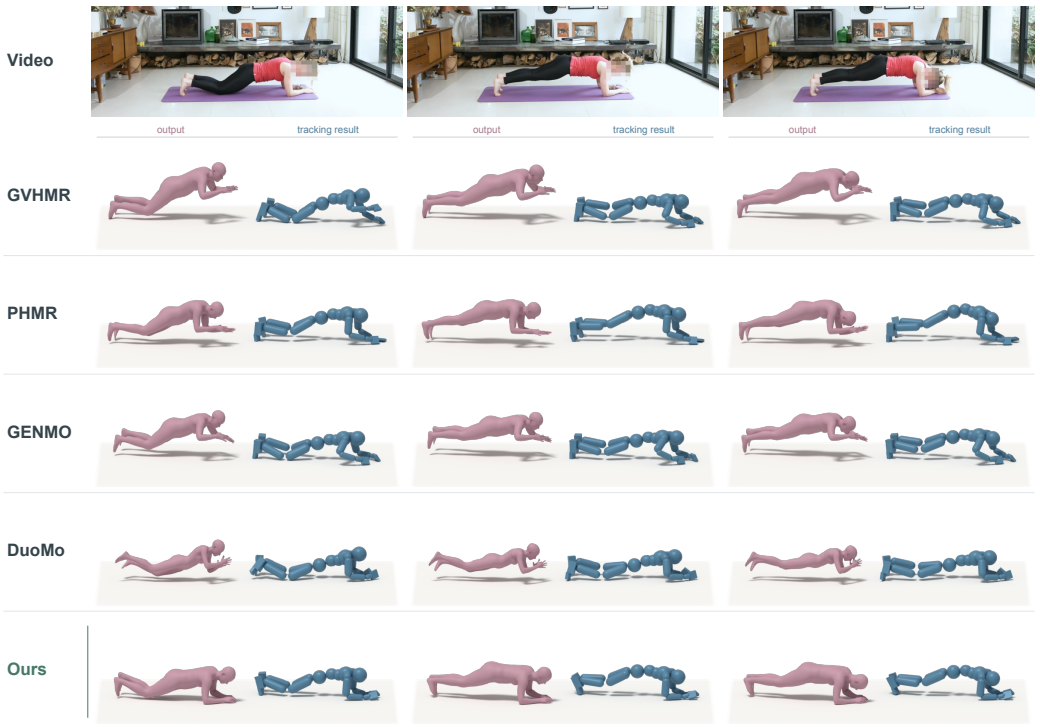



Figure D.2: **Qualitative comparison.** The top row shows the input video, followed by GVHMR, PromptHMR (PHMR), GENMO, DuoMo, and Ours.

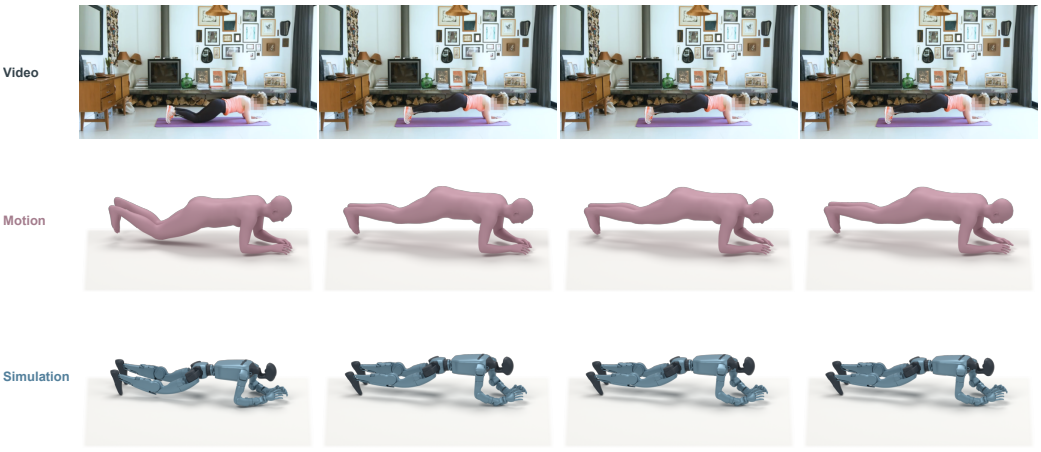



Figure D.3: **SONIC execution on Unitree G1.** Rows show the input video, the motion reconstructed by FlowHMR (pink), and SONIC tracking in Isaac Lab (blue).