PhysHMR: Learning Humanoid Control Policies from Vision for Physically Plausible Human Motion Reconstruction

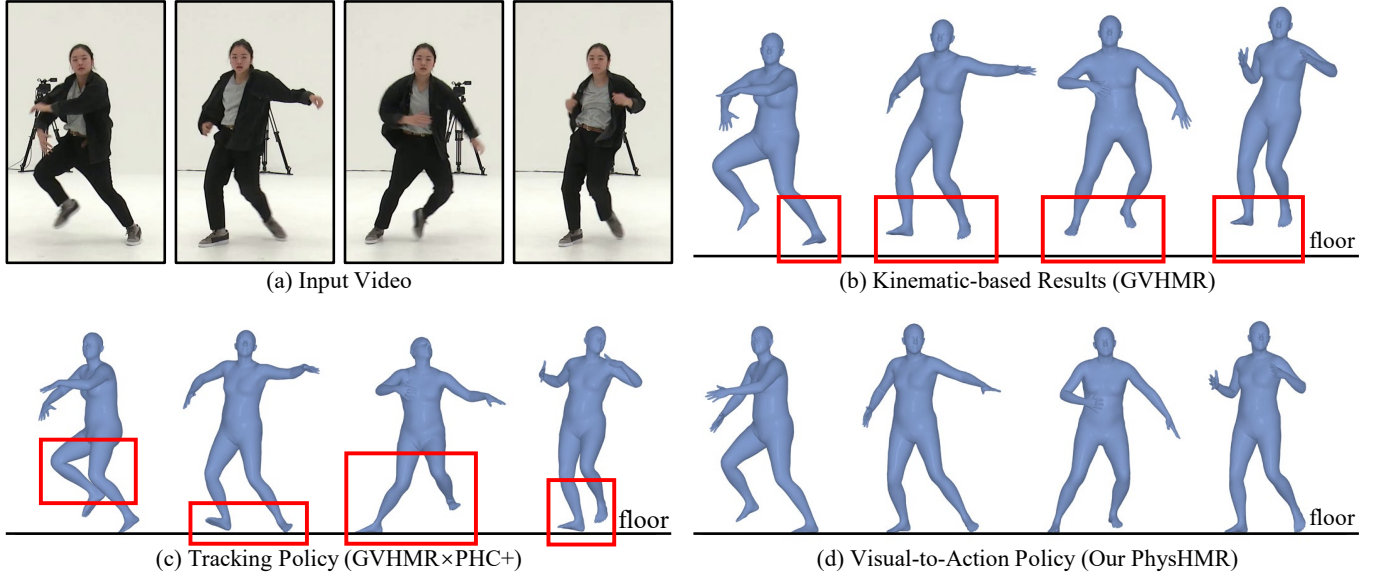QIAO FENG, YIMING HUANG, YUFU WANG, JIATAO GU, and LINGJIE LIU, University of Pennsylvania, USA

Fig. 1. Given a monocular video (a), (b) kinematic-based methods (e.g., GVHMR [Shen et al. 2024]) often cannot produce physically plausible results and suffer from artifacts like foot floating. (c) While tracking-based controllers (e.g., PHC+ [Luo et al. 2023]) can enforce physical plausibility, they may amplify errors from inaccurate motion reconstruction, leading to unnatural behaviors. (d) In contrast, our PhysHMR model learns a visual-to-action policy that directly predicts control signals from visual input, preventing error amplification and producing motions that are both physically plausible and visually aligned with the input video (a). As videos are the most effective way to assess the physical plausibility of the results, we encourage readers to view our suppl. video.

Reconstructing physically plausible human motion from monocular videos remains a challenging problem in computer vision and graphics. Existing methods primarily focus on kinematics-based pose estimation, often leading to unrealistic results due to the lack of physical constraints. To address such artifacts, prior methods have typically relied on physics-based post-processing following the initial kinematics-based motion estimation. However, this two-stage design introduces error accumulation, ultimately limiting the overall reconstruction quality. In this paper, we present PhysHMR, a unified framework that directly learns a visual-to-action policy for humanoid control in a physics-based simulator, enabling motion reconstruction that is both physically grounded and visually aligned with the input video. A key component of our approach is the pixel-as-ray strategy, which lifts 2D key-points into 3D spatial rays and transforms them into global space, providing robust global pose guidance without depending on noisy 3D root predictions. This soft global grounding, combined with local visual features from a pretrained encoder, allows the policy to reason over both detailed pose and global positioning. To overcome the sample inefficiency of reinforcement learning, we further introduce a distillation scheme that transfers motion knowledge from a mocap-trained expert to the vision-conditioned policy, which is then refined using physically motivated reinforcement learning rewards. Extensive experiments demonstrate that PhysHMR produces high-fidelity, physically plausible motion across diverse scenarios, outperforming prior approaches in both visual accuracy and physical realism.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Animation**.

Additional Key Words and Phrases: Motion reconstruction, Physical plausibility, Humanoid control, Monocular video

1 INTRODUCTION

Faithfully reconstructing human body dynamics from monocular videos, also known as Human Mesh Recovery (HMR), is a fundamental problem in computer vision, graphics, and robotics. Recent advances in human motion reconstruction [Rajasegaran et al. 2022; Shen et al. 2024; Shin et al. 2024; Sun et al. 2023; Wang et al. 2024a; Ye et al. 2023; Yuan et al. 2022] have achieved high accuracy in estimating body pose and shape. However, most existing methods overlook physically plausible body dynamics, leading to various artifacts such as foot sliding, ground penetration, and inconsistent contact behavior (see Fig. 1(b)). Achieving physically plausible human motion reconstruction remains an open and challenging problem.

Prior works have attempted to introduce physical constraints through a post-hoc correction stage. Some approaches incorporate analytical priors derived from rigid body dynamics, such as the

Authors' address: Qiao Feng, fengqiao@seas.upenn.edu; Yiming Huang, ymhuang9@seas.upenn.edu; Yufu Wang, yufu@seas.upenn.edu; Jiatao Gu, jgu32@cis.upenn.edu; Lingjie Liu, lingjie.liu@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Euler-Lagrange equation [Jiang et al. 2023; Zhang et al. 2024b,a], while others leverage reinforcement learning to train humanoid controllers that track pre-reconstructed motion [Yuan et al. 2021]. Although these methods improve physical realism to some extent, they share a *common limitation*: motion is first reconstructed from visual cues alone and then refined by a separate physics module. This decoupled design overlooks the ambiguity inherent in monocular videos, where multiple plausible motions can explain the same visual observation. Once a single solution is selected in the reconstruction stage, the downstream physics module can no longer access the full observational context, leading to suboptimal corrections and limited consistency with the visual evidence (see Fig. 1c).

In light of these limitations, we argue that a more effective approach is to **unify motion estimation and physical reasoning within a single framework**, allowing visual cues and physical constraints to inform the same decision process. To this end, we propose PhysHMR, a novel framework that directly learns a visual-to-action policy to control a simulated humanoid directly from monocular video observations, resulting in reconstructed motion that is both visually consistent and physically plausible. Unlike prior two-stage approaches, PhysHMR unifies the two stages through a single policy network that jointly reasons over visual observations and physical dynamics. By executing motion within a physics-based simulator [Makoviychuk et al. 2021], it naturally enforces physical constraints such as ground contact, joint limits, and momentum conservation. By conditioning the policy directly on image features, we can exploit rich visual context beyond skeletal pose estimations, enabling the humanoid to produce motion that faithfully aligns with the input video while adhering to physical laws.

Training high-dimensional visual-control policies purely with reinforcement learning is often sample-inefficient and unstable [Luo et al. 2024a]. To address these issues, PhysHMR proposes a distillation strategy that transfers knowledge from a mocap-trained imitation expert. Specifically, a pretrained visual encoder[Shen et al. 2024] extracts features from each video frame, which serve as local pose references for the control policy. These features retain rich pose information without committing to potentially inaccurate 3D reconstructions. The expert controller, trained on high-quality motion capture data, provides action supervision that imparts strong human motion priors, which significantly accelerates convergence and stabilizes learning. The policy is further refined with reinforcement learning, using a composite reward that balances motion imitation, realism through adversarial motion priors, and physical smoothness.

Since physical plausibility must be assessed in the global pose space rather than the local pose space, it is necessary to estimate global pose information (i.e., the root joint position) in addition to local pose references from images. However, predicting the 3D root joint position from monocular video is often noisy, which significantly compromises the robustness of policy learning. This is because inconsistencies between local pose estimates and erroneous 3D root predictions can lead to unnatural global motions—for example, the local pose may indicate forward movement, while the noisy root prediction pulls the motion backward, resulting in jittery or unstable behavior. Such mismatches make it difficult for the policy to learn physically consistent dynamics in global space. To address this, instead of relying on explicit 3D root prediction, we lift multiple detected 2D keypoints into 3D rays, which serve as a soft global pose reference. These spatial rays guide the policy in predicting actions that transform the humanoid into globally consistent poses without requiring strict absolute 3D supervision. This approach provides gentle global grounding, improves the robustness of policy learning, and enables physically plausible human motion reconstruction.

We evaluate PhysHMR on challenging motion datasets, including Human3.6M, AIST++, and EMDB2, showing comparable motion accuracy to state-of-the-art kinematics-based methods while significantly improving physical plausibility. Our approach reduces common non-physical artifacts (e.g. foot sliding, ground penetration), improving the suitability of reconstructed motion for downstream applications such as simulation, animation, and robotics.

In summary, our contributions are three-fold:

- We present PhysHMR, the **first** unified framework for jointly performing human motion perception and control, enabling high-quality and physically plausible human motion reconstruction from monocular videos.
- We introduce a novel distillation approach to distill a visual-to-action policy from a pretrained mocap imitation policy, which accelerates convergence and stabilizes policy learning.
- We propose a soft global grounding strategy by lifting 2D keypoints into 3D spatial rays, avoiding the need for noisy 3D root predictions and enabling robust policy learning of physically plausible motion in global space.

2 RELATED WORKS

2.1 Kinematics-Based Human Mesh Recovery

Parametric human models [Loper et al. 2015; Osman et al. 2020; Pavlakos et al. 2019; Xu et al. 2020] have been widely adopted to reconstruct human motion from monocular video. Early works [Arnab et al. 2019; Bogo et al. 2016; Huang et al. 2017; Xiang et al. 2019] focus on fitting these models to individual image frames. More recently, regression-based approaches, which leverage large-scale datasets, have gained attention for their ability to achieve general-purpose human mesh recovery [Cai et al. 2023; Goel et al. 2023; Yin et al. 2025]. To account for dynamic camera movements, data-driven methods have been extended to estimate per-frame camera poses [Shin et al. 2024; Sun et al. 2023; Yuan et al. 2022]. Additionally, SLAM (Simultaneous Localization and Mapping) techniques have proven effective for robust camera motion estimation, further enhancing human motion recovery in complex scenarios [Wang et al. 2024a]. Despite these advances in human mesh recovery, purely kinematic methods often exhibit artifacts like foot sliding, ground penetration, and momentum inconsistency.

To address such artifacts, prior works have used physical priors as an auxiliary supervision to encourage plausible dynamics. PhysPT [Zhang et al. 2024b] proposes a neural module that refines kinematic motion using differentiable Euler-Lagrange losses to enforce rigid-body dynamics. D&D [Li et al. 2022a] refines kinematic motion by estimating external forces and applying analytical physical computation to enforce consistency with Newtonian dynamics. While these methods improve physical realism to some extent, they operate as post-hoc refinement on kinematic reconstructions, making it difficult to recover from the ambiguity in the
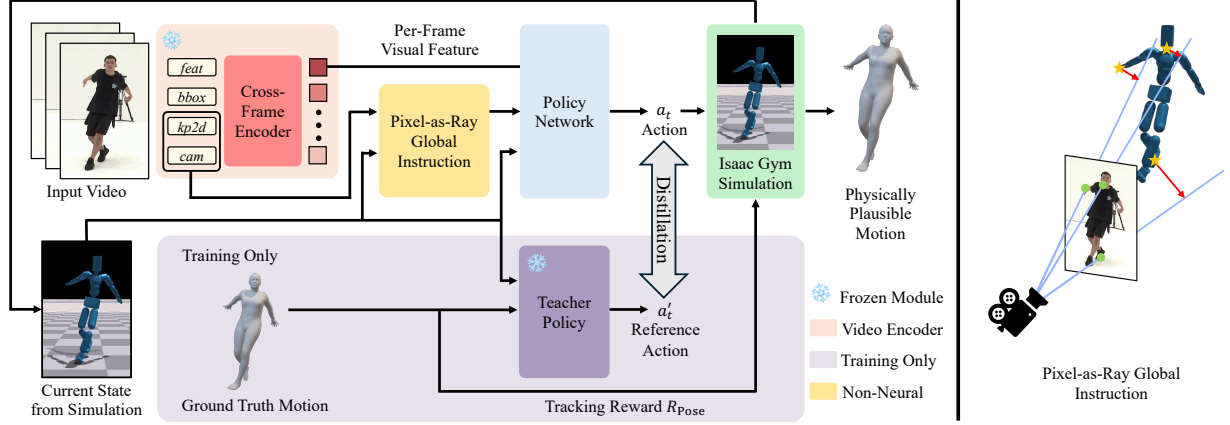
Fig. 2. Overview of our pipeline. A visual-to-action policy reconstructs physically plausible motion from monocular videos. Training efficiency is improved by combining reinforcement learning and knowledge distillation. Global motion is guided using a *pixel-as-ray* module that lifts 2D keypoints into 3D rays.

kinematics-based human mesh recovery stage. Moreover, the physical consistency is enforced through neural approximations rather than explicit physical simulation, leaving the overall pipeline fundamentally kinematics-based and decoupled from physical control.

2.2 Physics-based Human Motion Imitation

Physics simulation platforms [Makoviychuk et al. 2021; Todorov et al. 2012], combined with reinforcement learning, have enabled physically grounded control of simulated characters, producing highly realistic human motion [Dou et al. 2023; Peng et al. 2018, 2022, 2021a; Tessler et al. 2023; Wang et al. 2024b]. By training policies on large-scale motion capture datasets [Kobayashi et al. 2023; Mahmood et al. 2019; Peng et al. 2021c], many works have demonstrated high-fidelity motion imitation through learned control policies [Luo et al. 2024b, 2023, 2022; Tessler et al. 2024; Wagener et al. 2022]. However, these policies are trained to track clean 3D motion references, struggling to generalize when such data is unavailable. PHC [Luo et al. 2023] estimates 3D keypoints from video as motion references, but its two-stage design decouples control from visual input, often leading to jitter and unnatural motion.

Moreover, prior methods rely heavily on reinforcement learning, which typically suffers from low sample efficiency. Hence, they struggle to fully exploit rich visual information and instead depend primarily on sparse, deterministic inputs such as 3D keypoints or kinematics-based representations. simXR [Luo et al. 2024a] uses behavior cloning to learn control of simulated characters directly from images captured by head-mounted cameras, but lacks robustness due to limited data and lack of exploration. Learning vision-conditioned policies for human motion reconstruction that directly aligns with visual evidence remains a largely underexplored challenge.

3 PRELIMINARIES

We formulate physically plausible human motion reconstruction as a goal-conditioned, physics-based motion imitation problem. Specifically, we use deep reinforcement learning (DRL) to train a policy that drives a simulated humanoid [Luo et al. 2023] to imitate motion sequences within a physical environment, with the goal signals as

guidance. The policy, π, is modeled as a Markov Decision Process (MDP), defined by the tuple $\mathcal{M} = \langle S, A, T, R, \gamma \rangle$, where S, A, T, R, and γ denote the state space, action space, transition dynamics, reward function, and discount factor, respectively.

At each timestep t, the state s_t consists of proprioceptive information s_t^p and goal information s_t^g. Here, s_t^p includes the local 3D pose q_t and velocity $\dot{q}_t$. In traditional motion imitation tasks, the goal s_t^g is typically defined by a reference trajectory $(\theta_t, \Gamma_t, \tau_t)$, encoding the local pose, global translation, and global rotation. In our method, the goal information is extracted from the input video, including frame-level visual features and global spatial guidance computed by a pixel-as-ray strategy. This design enables the policy to directly leverage visual observations for motion imitation (see Sec. 4).

The action, a_t, specifies target joint rotations, which are provided as control targets to a proportional-derivative (PD) controller to generate physically valid motion. At each timestep t, the humanoid agent samples an action $a_t \in A$ from the policy $\pi(a_t|s_t)$, where $s_t \in S$ is the current state of the humanoid. The action is then executed in a physics simulator, producing the next state $s_{t+1} = T(s_t, a_t)$ and the reward $r_t = R(s_t, a_t)$ for this action. We optimize the policy using Proximal Policy Optimization (PPO), with the objective of maximizing the expected discounted return: $\mathbb{E}\left[\sum_{t=1}^{N} \gamma^{t-1} r_t\right]$.

4 METHOD

Fig. 2 provides an overview of our method. Given a monocular video with N frames $\{I_t\}_{t=1}^{N}$, our goal is to reconstruct a physically plausible human motion sequence, which consists of local poses $\{\theta_t \in \mathbb{R}^{23 \times 3}\}_{t=1}^{N}$, global translation $\{\Gamma_t \in \mathbb{R}^3\}_{t=1}^{N}$, and orientation $\{\tau_t \in \mathbb{R}^3\}_{t=1}^{N}$ in the world.

Our visual-to-action policy, PhysHMR, starts by extracting local visual features for each video frame using a pre-trained HMR model [Shen et al. 2024]. These features are then used as input to the policy network, which predicts control actions for the humanoid (Sec. 4.1). To provide spatial grounding, we propose a pixel-as-ray strategy that lifts 2D keypoints into 3D rays and transforms them into global space using camera poses. These global rays provide soft

global guidance to the policy without enforcing strict positional constraints (Sec. 4.2). Finally, we enhance the training of our visual-to-action policy through knowledge distillation from a pre-trained motion imitation expert, leading to improved sample efficiency and policy robustness (Sec. 4.3).

4.1 Local Reference from Visual Observations

Prior works [Shen et al. 2024; Ye et al. 2023] show that local motion features that capture relative joint articulation while being invariant to global root transformations are critical for effective motion learning. Such features are difficult to infer explicitly from images due to camera motion and depth ambiguity. Hence, we propose to use the pretrained video encoder from GVHMR [Shen et al. 2024], which is trained to predict SMPL joint rotations relative to their parent joints. This naturally yields root-invariant visual features that serve as structured and physically meaningful input for local control. Unlike explicit pose reconstructions that commit to a single, potentially inaccurate estimate, these visual features retain rich pose-related information without collapsing to a deterministic pose.

Given video frames $\{I_t\}_{t=1}^{N}$, we first preprocess each frame I_t to extract image features [Goel et al. 2023], bounding boxes [Jocher et al. 2023; Li et al. 2022b], 2D keypoints [Xu et al. 2022], and relative camera rotations [Teed et al. 2023], denoted as $f_t^{\text{feat}}, f_t^{\text{bbox}}, f_t^{\text{kp2d}}, f_t^{\text{cam}}$, respectively. These per-frame features are then fed into the video encoder, which aggregates information across frames:

$$\{F_t\}_{t=1}^{N} = \text{Enc}_{\text{GVHMR}}(\{f_t^{\text{feat}}, f_t^{\text{bbox}}, f_t^{\text{kp2d}}, f_t^{\text{cam}}\}_{t=1}^{N}) \in \mathbb{R}^{N \times D},$$

where D is the feature dimension. The cross-frame fusion process not only enhances stability under occlusion but also supports flexible feature masking, allowing the model to operate with partial inputs. This flexibility enables the use of motion-only datasets like AMASS [Mahmood et al. 2019] during training, even without paired RGB images, i.e., f_t^{feat} is dropped.

Although local features bolster reconstruction robustness, accurate physics learning still demands global guidance, because simulations must be carried out in the world coordinate frame. Thus, we enhance the local observation by leveraging GVHMR's multi-task MLP head to explicitly regress the future root orientation $\bar{\tau}_{t+1}$ from the visual features F_t. This auxiliary prediction provides a forward-looking estimate of the global root orientation in the camera coordinate system. We transform $\bar{\tau}_{t+1}$ into the world frame and compute its relative difference from the current root pose τ_t of the humanoid agent as:

$$\Delta \tau_t = \tau_t^{-1} \bar{\tau}_{t+1}.$$

This signal provides an explicit orientation cue that guides the agent's future heading. We include both the visual feature F_t and the relative root orientation $\Delta \tau_t$ in the observation passed to the policy at each timestep t.

4.2 Global Guidance via Pixel-as-Ray

Accurate global positioning is critical for physically plausible motion reconstruction, especially when camera motion is involved. However, directly predicting 3D trajectories from a monocular video is often unreliable due to depth ambiguity and motion noise. These trajectory errors can significantly degrade the performance of tracking-based control policies, leading to unstable motion. To circumvent this issue, we propose a pixel-as-ray strategy that encodes global guidance without enforcing explicit positional targets.

4.2.1 Keypoint Lifting to 3D Rays. Given extracted 2D keypoints, $f_t^{\text{kp2d}} = \{(u_t^i, v_t^i)\}_{i=1}^{J}$, that represent the image-space locations of each joint i of the simulated humanoid in frame t, and the camera intrinsics matrix $\mathbf{K}$, we back-project each keypoint to obtain a 3D ray in the camera coordinate system:

$$\text{ray}_t^i(s) = \mathbf{o}_t + s \cdot \mathbf{r}_t^i, \quad s > 0, \tag{1}$$

$$o_t = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{r}_t^i = \mathbf{K}^{-1} \begin{bmatrix} u_t^i \\ v_t^i \\ 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{o}_t$ is the camera origin in frame t, and $\mathbf{r}_t^i$ is the viewing direction for keypoint i. This ray represents all possible 3D positions of joint i along the corresponding viewing direction.

To align the ray with the simulation world, we further transform it using the camera-to-world transformation $\mathbf{T}_t^{\text{c2w}}$, which is estimated via off-the-shelf methods [Shen et al. 2024; Wang et al. 2024a]:

$$\hat{\mathbf{o}}_t = \mathbf{T}_t^{\text{c2w}} \cdot \mathbf{o}_t, \quad \hat{\mathbf{r}}_t^i = \mathbf{T}_t^{\text{c2w}} \cdot \mathbf{r}_t^i, \tag{3}$$

where $\hat{\mathbf{o}}_t$ denotes the ray origin in world coordinates, and $\hat{\mathbf{r}}_t^i$ is the transformed ray direction for keypoint i.

4.2.2 Computing Ray Displacement Vectors. For each humanoid joint i at time t, we compute the shortest vector from the joint position $\mathbf{j}_t^i$ to the corresponding ray defined by origin $\hat{\mathbf{o}}_t$ and direction $\hat{\mathbf{r}}_t^i$, yielding a displacement vector $\mathbf{d}_t^i$:

$$\mathbf{d}_t^i = \text{proj}_{\perp}(\mathbf{j}_t^i, \hat{\mathbf{o}}_t, \hat{\mathbf{r}}_t^i), \tag{4}$$

where $\text{proj}_{\perp}(\cdot)$ denotes the perpendicular offset from the point $\mathbf{j}_t^i$ to the ray $\hat{\mathbf{o}}_t + s \cdot \hat{\mathbf{r}}_t^i$, and $\mathbf{j}_t^i$ is obtained from the simulated humanoid proprioception s_t^p. These displacements $\{\mathbf{d}_t^i\}_{i=1}^{J}$ are concatenated and passed to the policy network as global spatial observations. Compared to using noisy 3D joint positions, this formulation enables more flexible and robust spatial grounding for humanoid control.

To account for potentially unreliable 2D keypoint estimates, we append the keypoint confidence scores to the displacement vectors, enabling the policy to adaptively modulate its reliance on uncertain inputs. Additionally, inspired by [Goel et al. 2023], we introduce random masking and perturbation of keypoint inputs during training to improve robustness under in-the-wild conditions.

4.3 Policy Learning with Reinforcement and Distillation

4.3.1 Distillation. Although we've used a pretrained visual encoder to extract informative features from monocular images, directly training a control policy from these features using reinforcement learning remains highly sample-inefficient. To address this, we introduce a knowledge distillation framework that transfers motion expertise from a pretrained teacher policy, $\pi_{\text{teach}}(a_t \mid s_t^p, \theta_t)$, which is trained to perform standard motion imitation using ground-truth pose supervision from the AMASS dataset [Mahmood et al. 2019]. The teacher policy takes as input the agent's proprioceptive state

s_t^p and the target kinematic pose θ_t, and outputs physically valid actions that track the reference motion.

Given paired supervision data (I_t, θ_t), where I_t is the input image and θ_t is the corresponding target kinematic pose, we use the teacher's action as a supervision signal to guide the training of our visual-to-action policy, $\pi_{\text{PhysHMR}}(a_t \mid s_t^p, F_t, \Delta\tau_t, \{\mathbf{d}_t^i\}_{i=1}^J)$. This policy takes as input the agent's proprioceptive state s_t^p, frame-level visual features F_t, relative root orientation $\Delta\tau_t$, and pixel-to-ray spatial displacements $\{\mathbf{d}_t^i\}_{i=1}^J$, and is trained to imitate the teacher's actions without using ground-truth pose targets as explicit input. The training objective minimizes a distillation loss between the actions predicted by the student and teacher policies:

$$L_{\text{distill}} = \left\| \pi_{\text{PhysHMR}}(a_t \mid s_t^p, F_t, \Delta\tau_t, \{\mathbf{d}_t^i\}_i) - \pi_{\text{teach}}(a_t \mid s_t^p, \theta_t) \right\|^2$$

Note that both policies operate on the same proprioceptive state s_t^p obtained from simulation, but differ in how goal information is provided: the teacher policy is conditioned on the explicit ground-truth pose θ_t, while the PhysHMR policy relies on features extracted from the input image. This objective encourages our PhysHMR policy to learn a control strategy capable of effectively imitating motion depicted in the input video without explicit pose signals.

4.3.2 Overall Loss. While distillation enables efficient learning from a strong teacher, it is inherently limited by the accuracy and diversity of the teacher's actions. To enable more accurate and adaptable motion control that goes beyond the fixed supervision provided by the teacher, we complement supervised knowledge distillation with reinforcement learning, allowing the PhysHMR policy to refine its behavior through dynamic interaction with the environment. Specifically, we utilize a composite reward function:

$$R(s_t^p, \theta_t) = \alpha_1 R_{\text{pose}} + \alpha_2 R_{\text{amp}} + \alpha_3 R_{\text{energy}}, \tag{5}$$

where θ_t is the target reference pose. R_{pose} is an imitation reward that promotes alignment with the reference pose, θ_t [Luo et al. 2023]. R_{amp} is a style-based reward via Adversarial Motion Priors (AMP) [Peng et al. 2021b] that encourages the generation of realistic, human-like motion aligned with the motion prior. R_{energy} is an energy reward that penalizes excessive joint accelerations to improve smoothness and reduce jitter [Winkler et al. 2022]. Additional details about the reward formulation are provided in the suppl. document.

The overall training objective combines both supervised and reinforcement signals:

$$L = L_{\text{distill}} + L_{\text{PPO}}, \tag{6}$$

where L_{PPO} is the loss term calculated by PPO using the reward in Eq. 5. By jointly training with distillation and reinforcement learning, our visual-to-action policy benefits from both sample-efficient supervision and environment-driven refinement, resulting in not only accelerated convergence but also enhanced policy performance.

5 EXPERIMENTS

5.1 Implementation Details

We use Isaac Gym [Makoviychuk et al. 2021] as the physics simulator and train our model on a single NVIDIA L40 GPU. The physics simulation runs at 60 Hz, while control actions are issued at 30 Hz.

All videos and motion sequences are sampled at 30 FPS for consistency. We parallelize training with 1,536 environments to improve sample efficiency. The main policy network is implemented as an MLP with hidden layer dimensions of [2048, 1536, 1024, 1024, 512, 512] and SiLU as the activation function. Reinforcement learning is conducted with Proximal Policy Optimization (PPO), using a clip coefficient of 0.2. PHC+[Luo et al. 2024b] is used as the teacher policy. The distillation loss is jointly optimized with the PPO objective. We apply gradient clipping with a threshold of 50 to ensure stability. Early termination is enabled to reduce ineffective exploration on failed episodes and accelerate convergence. The model typically converges after approximately three days of training.

Dealing with Shape Variance. Our approach does not rely on explicit shape information. We estimate the human shape parameters with an off-the-shelf tool and compute the scale difference relative to a zero-shape model. This scales the simulation space to match real-world units, such that the humanoid retains canonical zero-shape.

5.2 Datasets

We utilize Human3.6M [Ionescu et al. 2014], AIST++ [Li et al. 2021], EMDB2 [Kaufmann et al. 2023], and AMASS [Mahmood et al. 2019] for our experiments. Human3.6M contains 3.6M 3D human poses from 11 actors across 4 viewpoints, with accurate 3D keypoints but no SMPL ground truth. We exclude sequences involving chairs to avoid simulator inconsistencies. SMPL parameters are estimated via GVHMR and refined using LBFGS by aligning SMPL joints with the 3D keypoints. Note that since the full Human3.6M dataset is used in training of both TRAM and GVHMR, we only use it to conduct ablation studies. We use a zero-shape SMPL model and introduce an additional scale parameter to account for individual body proportions 5.1. AIST++ provides 1,408 dance sequences from 30 subjects across 9 views, featuring dynamic and diverse movements that are challenging for physics-based motion learning. AIST++ provides only a scale parameter without explicit shape information. AMASS is a large-scale, image-free motion capture dataset. EMDB2 contains long-range, moving-camera sequences. We remove sequences like skateboarding and stair climbing that are incompatible with simulation. We train PhysHMR on the combined training splits of Human3.6M, AIST++, and AMASS (image-free). Only AMASS is used for the AMP reward. EMDB2 is used for evaluation only.

5.3 Metrics

To evaluate the accuracy and physical plausibility of the reconstructed motion, we use the following metrics: (1) **MPJ** (Mean Per Joint Position Error, MPJPE, mm): Measures the average Euclidean distance between predicted and ground-truth 3D joint positions after aligning the root joint. (2) **WA** (World-aware MPJPE, WA-MPJPE, mm): Similar to MPJPE, but computed in the global coordinate system, capturing errors in both pose and global translation. (3) **PA** (Procrustes Aligned MPJPE, PA-MPJPE, mm): Computes MPJPE after applying rigid alignment (scale, rotation, translation) to isolate pose errors independent of global position. (4) **VEL** (Velocity Error, mm/s) and (5) **ACC** (Acceleration Error, mm/s^2): Measure the temporal consistency of joint movement across frames.
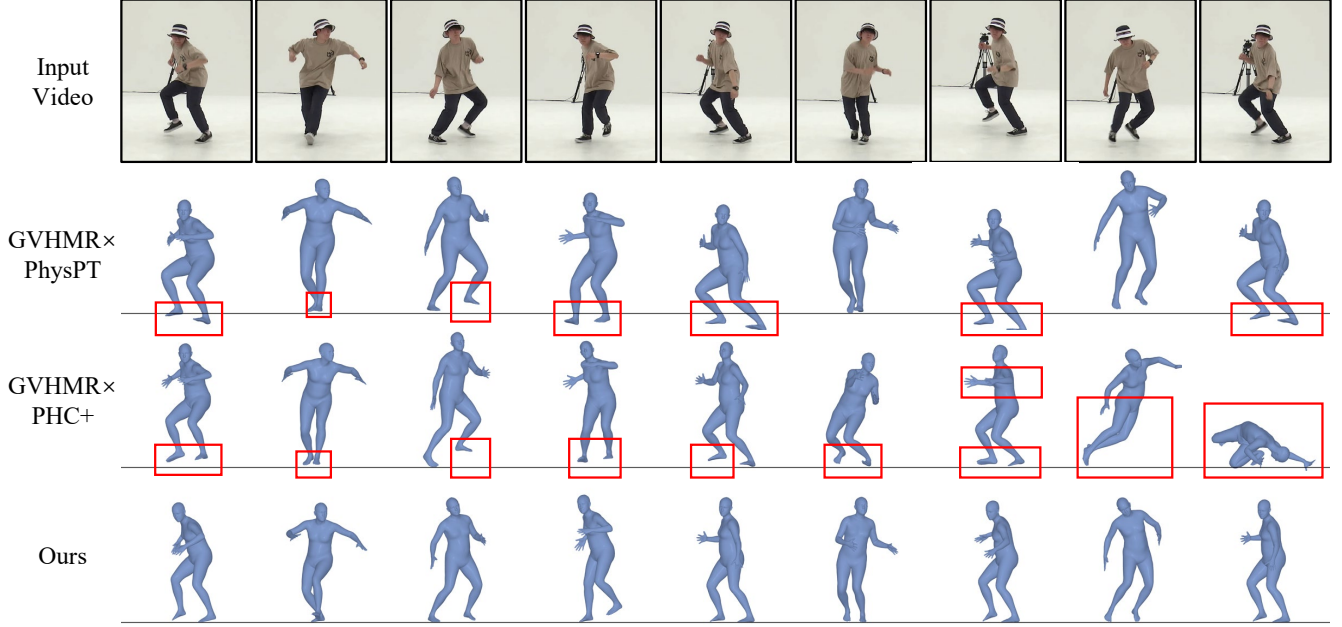
Fig. 3. Comparison against two physics-based methods. The black line indicates the ground. PhysPT (row 2) uses neural networks to approximate physics, but still suffers from ground penetration. PHC+ (row 3) amplifies motion reconstruction errors during tracking, leading to unstable results. Both methods cannot correct upstream errors. In contrast, our visual-to-action approach produces motion that is both physically plausible and visually aligned.

Table 1. Comparison of our motion reconstruction variants on AIST++ and EMDB2 under kinematic and physical plausibility metrics. Lower is better.

Phys. Type	Method	EMDB2							AIST++						
		PA	WA	MPJ	FS	HV	ACC	VEL	PA	WA	MPJ	FS	HV	ACC	VEL
Kin.	TRAM	**35.51**	**148.05**	<u>56.74</u>	11.76	22.97	**4.77**	**8.77**	50.18	189.44	<u>76.30</u>	23.18	7.93	9.31	21.85
	GVHMR	40.95	228.67	65.21	<u>5.65</u>	26.42	5.40	<u>10.19</u>	53.43	**175.64**	79.05	11.54	4.64	10.19	<u>14.40</u>
Neural	PhysPT (CLIFF)	48.40	762.78	77.00	11.02	6.54	6.72	19.92	70.72	260.07	108.57	13.68	3.71	9.21	17.05
	TRAM × PhysPT	39.90	704.57	61.42	8.49	7.02	5.38	17.71	52.79	250.30	83.93	<u>10.94</u>	3.55	<u>8.59</u>	16.00
	GVHMR × PhysPT	41.34	682.03	66.08	10.71	8.46	<u>5.35</u>	17.24	55.29	235.66	83.93	11.24	<u>3.46</u>	8.90	15.36
Track.	TRAM × PHC+	52.94	<u>158.58</u>	74.34	23.41	7.64	9.56	14.00	71.21	212.28	101.70	36.57	4.95	12.55	22.74
	GVHMR × PHC+	46.24	193.01	72.50	12.71	7.71	7.43	12.21	67.38	193.23	109.77	24.79	6.05	10.05	17.10
V2A	**Ours**	<u>39.34</u>	189.26	**55.48**	**4.60**	**5.04**	5.49	10.53	<u>50.40</u>	<u>187.42</u>	**63.94**	**9.14**	**3.10**	**6.58**	**12.13**

To assess physical realism, we introduce a new metric: (6) **HV** (Foot Height Variance, mm): In every frame, we record the vertical position of the lowest foot joint. We select the lowest 25 % across all frames and compute their variance; smaller HV indicates more stable, physically realistic contact. This kinematic proxy evaluates contact consistency without requiring an explicit ground plane. Additionally, we use (7) **FS** (Foot Sliding, mm) to measure undesired foot movement when the foot is expected to be in contact with the ground. Together, these metrics provide a comprehensive evaluation of both motion accuracy and physical plausibility.

5.4 Comparisons

We compare our method with both kinematic and physics-based state-of-the-art approaches. Kinematic methods, TRAM [Wang et al. 2024a] and GVHMR [Shen et al. 2024], estimate human motion from videos without enforcing physical constraints. In contrast, PhysPT [Zhang et al. 2024b] introduces a physics-based approach by first estimating SMPL parameters using CLIFF [Li et al. 2022b] and then refining the motion with a transformer-based model to improve physical plausibility. We also provide results for GVHMR × PhysPT and TRAM × PhysPT, where the SMPL estimation backbone of PhysPT is replaced with TRAM and GVHMR, respectively, to ensure a fair comparison. Additionally, we evaluate tracking-based methods, TRAM × PHC+ and GVHMR × PHC+, where the tracking policy PHC+ [Luo et al. 2024b] is applied to track the outputs of TRAM and GVHMR, respectively, providing a direct comparison between motion reconstruction via traditional tracking policies and our visual-to-action policy.

5.4.1 Quantitative Results. As shown in Tab. 1, kinematic-based methods generally achieve lower errors on MPJPE, PA-MPJPE, and WA-MPJPE, as they are directly optimized to minimize 3D keypoint discrepancies and ignore physical constraints. In contrast, physics-based approaches trade off keypoint accuracy for physical plausibility. For example, PhysPT, TRAM × PhysPT, and GVHMR × PhysPT all achieve better physical metrics due to PhysPT's physics-aware design. However, its global trajectory relies on foot-ground contact prediction, which can be inaccurate and result in high WA-MPJPE, especially for long-range motions in EMDB2.

Traditional tracking-based methods, TRAM × PHC+ and GVHMR × PHC+, exhibit stable performance when the kinematic estimates are accurate, but their quality degrades severely when those estimates are poor, as seen on AIST++, where the challenging motions lead to high PA-MPJPE. Moreover, such methods fails to capitalize on physical simulation, with subpar FS and HV scores. This is due to excessive movements of the limbs during balance recovery, which degrades physical metrics.

In contrast, our method achieves state-of-the-art performance on FS and HV, demonstrating superior physical realism. It also remains competitive across MPJPE metrics. By learning policies directly from visual features, our approach can produce 3D human motion that is both physically plausible and visually aligned with the input video.

5.5 Ablation Study

We conduct ablation experiments on H36M to validate the effectiveness of our proposed *pixel-as-ray* formulation and the combined training strategy based on distillation and reinforcement learning. **Effect of Pixel-as-Ray.** Table 2 evaluates the impact of different global instruction strategies. Removing the global instruction entirely (ImgFeat) yields good PA-MPJPE and MPJPE, but significantly worse WA-MPJPE, indicating that the humanoid mimics local motions well but fails to track global trajectories. Replacing pixel-as-ray with global supervision from explicit root-relative displacements estimated with GVHMR (+ 3D root) results in degraded performance across all metrics, as errors in root estimation introduce misleading guidance that conflicts with local motion. In contrast, using 2D keypoints via pixel-as-ray (+ pixelray) provides more robust and relaxed global instruction, achieving comparable PA-MPJPE to the no-global setting while substantially improving WA-MPJPE. **Effect of Distillation.** Table 3 and Figure 4 compare different training strategies. Using PPO Only leads to slow convergence and suboptimal final performance. Behavior cloning (i.e., Distillation Only) enables faster early-stage learning but lacks exploration, resulting in limited reward improvements. Note that in the Distillation Only setting, the reward is computed only for evaluation and not used during training. Our joint training strategy combines the strengths of both: it accelerates convergence and achieves higher final rewards, while also delivering better generalization on test sequences.

6 CONCLUSION

We presented PhysHMR, a unified framework for reconstructing physically plausible human motion from monocular videos by directly mapping visual inputs to humanoid control actions. Unlike

Table 2. Ablation on global instruction strategies.

Obs. Type	PA-MPJPE ↓	WA-MPJPE ↓	MPJPE ↓
ImgFeat	**35.85**	142.17	47.78
+ 3D root	38.70	195.59	65.03
+ pixelray	36.05	**112.60**	**47.01**

Table 3. Comparison of policy learning strategies. Combining reinforcement learning (PPO) and distillation yields the best performance.

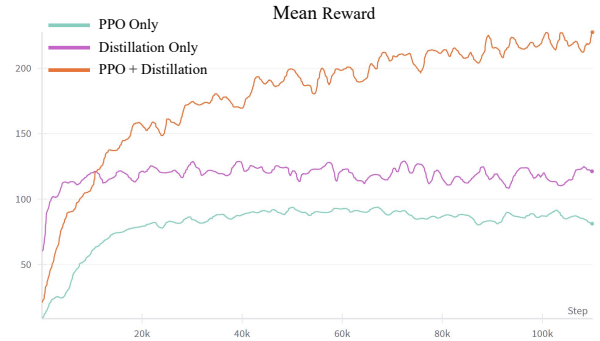| Training Strategy	PA-MPJPE ↓	WA-MPJPE ↓	MPJPE ↓
PPO Only	42.18	117.33	58.25
Distillation Only	39.62	114.69	52.41
PPO + Distillation	**36.05**	**112.60**	**47.01**

Fig. 4. Mean reward curves during training. PPO Only converges slowly and underperforms. Distillation Only converges quickly but plateaus early. Our approach (PPO + Distillation) achieves both faster convergence and higher final rewards.

prior methods, PhysHMR learns a visual-to-action policy that integrates physical dynamics during inference. To improve efficiency and robustness, we introduce motion distillation from a mocap-trained expert and a novel pixel-as-ray strategy that provides soft global guidance without relying on noisy 3D root predictions. **Limitation and Future Work.** While PhysHMR generates high-fidelity motion, a sim-to-real gap persists due to differences in body mechanics and contact properties. Future work will incorporate personalized physical parameters to better reflect real-world dynamics. Additionally, motion reconstruction from a single monocular video is underconstrained due to ambiguity and occlusion; using a conditional generative model instead of a deterministic policy may better capture diverse and physically plausible motions.

REFERENCES

Anurag* Arnab, Carl* Doersch, and Andrew Zisserman. 2019. Exploiting temporal context for 3D human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 561–578.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2023. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*.

Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. 2023. C·ASE: Learning Conditional Adversarial Skill Embeddings for Physics-based Characters. *SIGGRAPH Asia 2023 Conference Papers* (2023).

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *ICCV*.

Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. 2017. Towards Accurate Markerless Human Shape and Pose Estimation over Time. In *International Conference on 3D Vision (3DV)*.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

Yifeng Jiang, Jungdam Won, Yuting Ye, and C Karen Liu. 2023. Drop: Dynamics responses from human motion prior and projective dynamics. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. 2023. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*.

Makito Kobayashi, Chen-Chieh Liao, Keito Inoue, Sentaro Yojima, and Masafumi Takahashi. 2023. Motion Capture Dataset for Practical Use of AI-based Motion Editing and Stylization. arXiv:2306.08861 [cs.CV]

Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. 2022a. D&D: Learning Human Dynamics from Dynamic Camera. In *European Conference on Computer Vision*.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. arXiv:2101.08779 [cs.CV]

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. 2022b. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *ECCV*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Zhengyi Luo, Jinkun Cao, Rawal Khirodkar, Alexander Winkler, Kris Kitani, and Weipeng Xu. 2024a. Real-Time Simulated Avatar from Head-Mounted Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 571–581.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. 2024b. Universal Humanoid Motion Representations for Physics-Based Control. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=OrOd8PxOO2

Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. 2023. Perpetual Humanoid Control for Real-time Simulated Avatars. In *International Conference on Computer Vision (ICCV)*.

Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. 2022. Embodied Scene-aware Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *ICCV*.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. 2021. Isaac Gym: High Performance GPU Based Physics Simulation For Robot Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=fgFBtYgJQX_

Ahmed A A Osman, Timo Bolkart, and Michael J. Black. 2020. STAR: A Sparse Trained Articulated Human Body Regressor. In *European Conference on Computer Vision (ECCV)*. 598–613. https://star.is.tue.mpg.de

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021c. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. DeepMimic: Example-guided Deep Reinforcement Learning of Physics-based Character Skills. *ACM Trans. Graph.* 37, 4, Article 143 (July 2018), 14 pages. https://doi.org/10.1145/3197517.3201311

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. 2022. ASE: Large-scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters. *ACM Trans. Graph.* 41, 4, Article 94 (July 2022).

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. 2021a. AMP: adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.* 40, 4, Article 144 (July 2021), 20 pages. https://doi.org/10.1145/3450626.3459670

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. 2021b. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Trans. Graph.* 40, 4, Article 1 (July 2021), 15 pages. https://doi.org/10.1145/3450626.3459670

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. 2022. Tracking People by Predicting 3D Appearance, Location & Pose. In *CVPR*.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia Conference Proceedings*.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. 2024. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. In *Computer Vision and Pattern Recognition (CVPR)*.

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. 2023. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Zachary Teed, Lahav Lipson, and Jia Deng. 2023. Deep Patch Visual Odometry. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 39033–39051. https://proceedings.neurips.cc/paper_files/paper/2023/file/7ac484b0f1a1719ad5be9aa8c8455fbb-Paper-Conference.pdf

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. 2024. MaskedMimic: Unified Physics-Based Character Control Through Masked Motion Inpainting. *ACM Transactions on Graphics (TOG)* (2024).

Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. 2023. CALM: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–9.

Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5026–5033. https://doi.org/10.1109/IROS.2012.6386109

Nolan Wagener, Andrey Kolobov, Felipe Vieira Frujeri, Ricky Loynd, Ching-An Cheng, and Matthew Hausknecht. 2022. MoCapAct: A Multi-Task Dataset for Simulated Humanoid Control. In *Advances in Neural Information Processing Systems*, Vol. 35. 35418–35431.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. 2024a. TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos. In *European Conference on Computer Vision*. Springer, 467–487.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Ailing Zeng, Jing Lin, Zhengyi Luo, Hok Wai Tsui, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. 2024b. SkillMimic: Learning Reusable Basketball Skills from Demonstrations. arXiv:2408.15270v1 [cs.CV]

Alexander Winkler, Jungdam Won, and Yuting Ye. 2022. QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/3550469.3555411

Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*. 6184–6193. https://openaccess.thecvf.com/content_CVPR_2020/html/Xu_GHUM__GHUML_Generative_3D_Human_Shape_and_Articulated_Pose_CVPR_2020_paper.html

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2023. Decoupling Human and Camera Motion from Videos in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. 2025. SMPLest-X: Ultimate Scaling for Expressive Human Pose and Shape Estimation. *arXiv preprint arXiv:2501.09782* (2025).

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. 2022. GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. 2021. SimPoE: Simulated Character Control for 3D Human Pose Estimation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. 2024b. 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.* 2305–2317.

Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. 2024a. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 6164–6174.

A ABOUT CAMERA-TO-WORLD TRANSFORMATION

The key to linking the simulated world with the image domain is the T_t^{c2w} transformation. There are multiple ways to compute it, and our method does not depend on a specific solution.

In the EMDB2 experiments, we combine TRAM and GVHMR. TRAM provides accurate SLAM-based camera trajectories, but its result may differ from the global frame by a rigid transformation. GVHMR, on the other hand, does not model the camera explicitly but produces outputs in gravity-aligned space, where the gravity direction is downward.

For moving camera, we ignore translation and compute a global rotation by aligning the joint positions of GVHMR and TRAM (e.g., using Procrustes alignment). Once the rotation is known, we determine the translation by aligning the first-frame SMPL root position. We define the floor based on the foot position of the GVHMR result in the first frame, which is also used to initialize the humanoid in the simulator, similar to the PHC series. Small alignment errors are acceptable in practice. While the estimated ground height may contain small errors, it is acceptable in practice.

For static cameras, the process is simpler: we directly align the GVHMR first-frame gravity space result with the camera space.

B DEALING WITH SHAPE VARIANCE

In practice, the shape of a real human SMPL model can be approximated as a scaled version of the zero-shape SMPL. However, the humanoid used in the simulator always follows the zero-shape size. To resolve this mismatch, we apply a scale correction to the transformation:

$$\hat{T}_t^{c2w} = T_t^{c2w} \cdot \frac{1}{\text{scale}}$$

C REWARD DEFINITION

Following PHC+, we use a SMPL-based humanoid agent, which consists of 24 rigid bodies, 23 of which are actuated. The proprioceptive state s_t^p is defined as:

$$s_t^p := (r_t, p_t, v_t, \omega_t) \tag{7}$$

where r_t, p_t, v_t, and ω_t are the simulated joint rotations, positions, velocities, and angular velocities, respectively. The reference state θ_t is defined as:

$$\theta_t := (\hat{r}_{t+1} \ominus r_t, \hat{p}_{t+1} - p_t, \hat{v}_{t+1} - v_t, \hat{\omega}_{t+1} - \omega_t, \hat{r}_{t+1}, \hat{p}_{t+1}) \tag{8}$$

where $\ominus$ denotes rotation difference. $\hat{r}_{t+1}, \hat{p}_{t+1}, \hat{v}_{t+1}$, and $\hat{\omega}_{t+1}$ represent the reference joint rotations, positions, velocities, and angular velocities, respectively. The imitation reward R_{pose} is defined as:

$$\begin{aligned} R_{\text{pose}} := & w_p e^{-\lambda_p \|p_t - \hat{p}_t\|} + w_r e^{-\lambda_r \|r_t - \hat{r}_t\|} \\ & + w_v e^{-\lambda_v \|v_t - \hat{v}_t\|} + w_\omega e^{-\lambda_\omega \|\omega_t - \hat{\omega}_t\|} \end{aligned} \tag{9}$$

where $w_{\{\cdot\}}, \lambda_{\{\cdot\}}$ denote the corresponding weights. We utilize the same weight settings as PHC+.

To obtain the style reward, R_{amp}, we train a discriminator $D(s_{t-10:t}^p)$ jointly with the policy network to distinguish real motion sequences from those generated by the policy. The discriminator produces a scalar value based on the proprioception of the humanoid, encouraging the generation of realistic, human-like motion aligned with the motion prior.

D TEST DATA

We exclude AIST++ sequences with high heels and ballet shoes, resulting in 326 videos from 4 camera views (totaling 1304 videos). From EMDB2, we remove sequences involving skateboards or stairs, which cannot be replicated in the simulator. The remaining sequences are:

P0_09, P2_19, P2_20, P2_24, P3_27, P3_28, P4_35, P4_36, P4_37, P5_40, P7_55, P7_61, P8_65, P9_79, P9_80

In P2_24, we removed frames 1650–1800 due to a step-up motion. The sequence was split into P2_24_0 and P2_24_1. For P4_36, we removed the first 300 frames, and for P7_61, the first 600 frames (sitting or lying).

E EVALUATION PROTOCOL

Physics-based methods are generally less stable. Once failure occurs, large errors dominate, making the overall average error meaningless. To address this, we split all test sequences into 100-frame clips and evaluate them individually. Following PHC+, we compute metrics only on successfully executed sequences, discarding those with large errors (e.g., PA-MPJPE > 100). Splitting into smaller segments avoids excessive sequence removal, while ensuring that the final metrics remain representative.