Improving Human Motion Plausibility with Body Momentum

Ha Linh Nguyen
hlinhn@comp.nus.edu.sg
Tze Ho Elden Tse
eldentse@comp.nus.edu.sg
Angela Yao
ayao@comp.nus.edu.sg
National University of Singapore, Singapore

Abstract

Many studies decompose human motion into local motion in a frame attached to the root joint and global motion of the root joint in the world frame, treating them separately. However, these two components are not independent. Global movement arises from interactions with the environment, which are, in turn, driven by changes in the body configuration. Motion models often fail to precisely capture this physical coupling between local and global dynamics, while deriving global trajectories from joint torques and external forces is computationally expensive and complex. To address these challenges, we propose using whole-body linear and angular momentum as a constraint to link local motion with global movement. Since momentum reflects the aggregate effect of joint-level dynamics on the body's movement through space, it provides a physically grounded way to relate local joint behavior to global displacement. Building on this insight, we introduce a new loss term that enforces consistency between the generated momentum profiles and those observed in ground-truth data. We evaluate our loss on the global motion recovery task. Incorporating our loss reduces foot sliding and jitter, improves balance, and preserves the accuracy of the recovered motion. Code and data are available at the project page.

1 Introduction

Accurately modeling global human motion is crucial for predicting and analyzing human activities, enabling advancements in applications such as robotics, virtual and augmented reality, and autonomous systems. This process requires not only realistic combinations of joint rotations - commonly emphasized in motion recovery and prediction tasks [38] - but also a plausible trajectory of the root joint. These two components are inherently connected: the center of mass (CoM) of the body moves in response to environmental interactions, which are influenced by changes in joint configurations. Plausible motions must maintain consistency between global trajectories and local joint movements. For instance, a sideways stepping motion should correspond to a lateral global trajectory rather than one indicative of forward or backward walking. Similarly, in a high jump or a back flip, due to the conservation of

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

angular momentum, drawing one's limbs closer while spinning should increase the global rotation speed, not decrease it.

However, many works treat these two components of human motion separately [13, 27] or leave their relationship to be inferred from data alone [24, 46], leading to incompatibilities that render the motion implausible. A common manifestation of mismatches between the global root trajectory and joint configurations is unexplained root movements. For instance, the root may translate without corresponding changes in contact states, causing unnatural sliding of the feet or other grounded body parts.

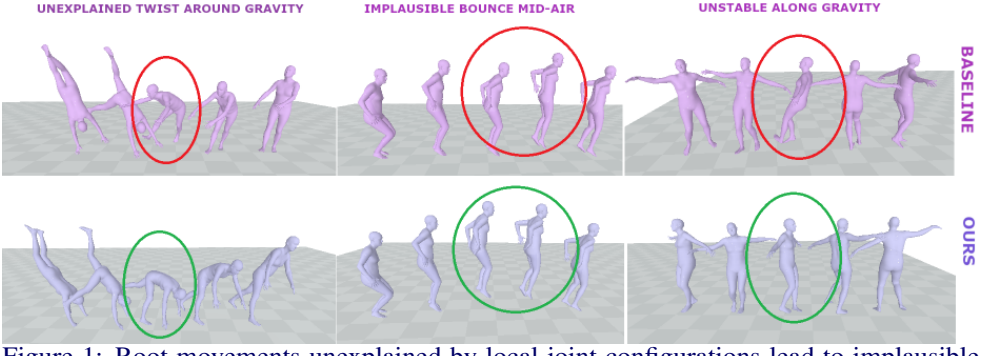

Figure 1: Root movements unexplained by local joint configurations lead to implausible motion (left to right: WHAM [50], PhysPT [72], GLAMR [67]). Our loss encourages more natural motion.

To address this issue, three main strategies have emerged: physics-based methods, learning-based approaches, and enhanced kinematic models. Physics-based methods use simulators to model human motion dynamics [48, 49, 68], but they rely on detailed environmental knowledge to compute interaction forces and timing. This often requires simplified representations of the body or surroundings [17, 48], limiting their effectiveness in complex, dynamic, or cluttered scenes. Moreover, optimizing contact forces and timing remains difficult due to their discrete nature. Learning-based methods attempt to predict these dynamic quantities or ease the optimization process [7, 28, 62, 72], yet they struggle with limited ground truth data. Annotations generated via inverse dynamics are often unreliable, as they rely on residual forces to correct modeling inaccuracies [60, 65]. Alternatively, some approaches improve the realism of purely kinematic motion by targeting specific elements crucial for plausibility - such as balance [55], ground interaction pattern [32], contact or friction [71]. While effective at reducing artifacts like foot sliding or interpenetration, they tend to focus on body-ground interactions and may overlook dynamics in aerial or acrobatic motions like those in basketball, parkour, or gymnastics.

During intervals between contact events, the body's rotation is entirely determined by the joint configurations, which influence its moment of inertia. Meanwhile, the translation of the CoM remains unaffected by changes in joint arrangements, as linear momentum is conserved. Even in more common cases with frequent ground contact where momentum is generally not conserved, there are discernible patterns in how momentum terms evolve over time [6, 16, 35, 39, 44]. Specifically, the angular momentum about the CoM is regulated to be close to 0 in a large class of movements [39, 43], while the control of linear momentum implies the control of the CoM trajectory, which tends to remain inside or close to the base of support for balance [33]. These patterns establish a relationship between joint movements as

observed in the root frame, and the global motion of the body.

From these observations and insights, we propose a novel loss term that provides an explicit connection between global movement to local joint configuration. Unlike prior physics-based methods that emphasize modeling forces and torques—which are often unobservable or difficult to estimate from real-world motion data—our approach leverages more accessible, yet still physically meaningful, features of human motion. Specifically, we guide the generation or reconstruction of motion by encouraging alignment with ground truth data in terms of whole-body linear and angular momentum, computed in a fixed world frame. These two global motion features capture the overall translation and rotation of the body in a way that is tightly coupled with local joint behaviors, and are critical for producing physically plausible and visually natural motion. By incorporating this loss into our training objective, we ensure that the resulting motions are not only kinematically valid but also globally consistent with how real bodies move in space.

Integrating our loss into existing methods leads to more plausible motion, as evidenced by reduced foot sliding, less unwanted jitter, and greater body stability, all without compromising performance metrics such as accuracy.

To summarize: (1) We propose a novel loss term based on momentum-related features of motion to improve consistency between global and local motion in reconstructed human motion. (2) We demonstrate that this term enhances the physical plausibility of motion and can be seamlessly integrated into a variety of existing models.

2 Related work

Kinematics-based methods. Motion models are learned models that encapsulate knowledge about how humans typically move. This knowledge can be leveraged to make future predictions [3, 66], fill in missing data [47, 53], or constrain solutions in tasks related to human motion. For example, many optimization-based methods for pose estimation or motion generation rely on motion models to provide feedback on how plausible a predicted or generated motion is [14, 22, 25]. Kinematics-based methods map motion sequences [5, 13, 24, 27, 40, 41, 70] or frame-to-frame transitions [15, 46] to the latent space of generative models using only kinematic features such as joint positions, rotations, or velocities. While effective for tasks like motion prediction in the root frame [36, 63] or reconstruction in the camera frame [2, 11], these models ignore the underlying dynamics of motion, often resulting in implausible global behaviors such as foot sliding or unrealistic root joint acceleration in mid-air. Our work complements these models by introducing a simple loss term that encourages physically consistent motion through global momentum features.

Physics-based methods. Recognizing the limitation of relying on kinematics alone, many methods have increasingly made use of physical laws in both regression-based and optimization-based approaches. Some methods use physics-related constraints to post-process generated or reconstructed motion samples, for example, by adhering more faithfully to contact events with the ground or estimating joint torques and ground reaction forces [8, 9, 45, 48, 62]. The common pipeline makes use of a physics simulator to advance the estimated kinematic pose to the next time instant given these constraints, denoising the kinematic estimation in the process [8, 48, 68, 69]. Post-processing is challenging and might distort the generated motion enough to lose its realism. An additional challenge is posed by the difficulty of integrating physics simulator in the pipeline, since contact events are usually discrete and, therefore, not amenable to easy inclusion in the usual machine learning model.

To address this challenge, recent works propose to replace the physics simulator with neural components that assume the same duties, such as contact detection and force calculation [28, 49, 72, 73]. They generally formulate the optimization problem with a smooth contact model to make the whole system differentiable. The estimated forces and contacts are encouraged to generate motions that match kinematics estimations, however the ground truth forces and torques are themselves generated from inverse dynamics and optimization procedure, which are sometimes unreliable due to modeling error [60, 65].

Another general approach avoids relying on unobservable quantities such as forces and torques by instead examining alternative markers of motion plausibility, for example balance [54, 55], ground interaction pattern [32], smoothness and friction [71]. While these methods are simpler and less computationally expensive than simulator-based counterparts, their heavy reliance on ground interaction patterns overlooks broader dynamical consistency. In contrast, our approach explicitly constrains momentum profiles, ensuring physically plausible movement across the entire body. Centroidal dynamics have been explored in past works for motion planning [26, 61], however they focus on trajectory optimization rather than directly enforcing momentum consistency in generated motions.

3 Methodology

3.1 Preliminaries

SMPL body partition. The center of mass of the entire body changes with different joint configurations, but each body part's center of mass remains approximately fixed relative to that body part as the body moves. A system's linear momentum and angular momentum are sums of the respective momenta of its parts. Therefore, we can calculate the momentum terms by dividing the body into parts and pre-compute some commonly needed quantities.

We partition the SMPL [31] body mesh into $P = 20$ parts. For each body part, we compute the volume V_i, the centroid c_i^0 and the moment of inertia I_i^0 of part i-th at the canonical pose. We normalize the mass by setting all subjects' total mass to 1. Assuming a uniform distribution of mass, the mass of each part m_i is then: $m_i = 1 \cdot \frac{V_i}{\sum_{i=1}^{P} V_i}$.

Variables definition. Let W be an inertial frame fixed to the world, its origin coincides with the root position and orientation at $t = 0$. We define frame B as the non-rotating frame that translates with the CoM while maintaining instantaneous alignment in orientation with the world frame W. All variables are in frame W unless otherwise indicated. $R(t)$ is the global rotation of the body; $T(t)$ is the global translation of the body; $\theta_i(R)$ is the rotation of body part i-th, $c_i^W(\theta_i, T)$ and $c_i^B(\theta_i)$ are the centroid position of body part i-th in frame W and frame B respectively, $I_i(\theta_i)$ is the moment of inertia of part i at t, where $I_i(\theta_i) = \theta_i I_i^0 \theta_i^{-1}$. Predicted quantities are denoted with a hat (e.g. $\hat{R}$ denotes the estimated global rotation) while their ground-truth counterparts are written without it. We omit time index for simplicity.

Body linear momentum. The linear momentum of the whole human body is the vector sum of the linear momenta of all its constituent parts and is therefore:

$$\mathrm{LMo}(R, T, \theta) = \sum_{i=1}^{P} m_i \dot{c}_i^W. \tag{1}$$

Body angular momentum. Similarly, the body angular momentum is given by:

$$\mathrm{AMo}(R, \theta) = \sum_{i=1}^{P} I_i \dot{\theta}_i + m_i c_i^B \times \dot{c}_i^B, \tag{2}$$

where the angular momentum of each body part is composed of the angular momentum about its own CoM, plus the angular momentum of its CoM about the chosen origin.

3.2 Frequency characteristics of momentum

The quasi-periodic nature of the spin angular momentum in walking [56] suggests that the frequency decomposition of this signal may hold information about the underlying dynamics. While the class of motions we investigate is broader and may lack the same quasi-periodic structure, we still anticipate that this signal in the frequency domain exhibits certain patterns.

Based on our analysis below, one identifiable characteristic of motion signal in frequency domain is the small magnitudes of its high-frequency components. Angular momentum L changes when there is a net external torque $\tau \neq 0$:

$$\tau = \frac{dL}{dt} \therefore \Delta L(t) = L_t - L_0 = \int_0^t \tau(t)dt = \int_{-\infty}^t \tau(t)dt.$$

Therefore, let $\mathcal{F}(f(t))$ be $f(t)$ in frequency domain and $\widehat{f}(w)$ be the complex number conveying the amplitude and phase of $\mathcal{F}(f(t))$ at frequency w,

$$\mathcal{F}(\Delta L(t)) = \mathcal{F}\left(\int_{-\infty}^t \tau(t)dt\right) \implies \widehat{\Delta L}(w) = \frac{1}{jw}\widehat{\tau}(w).$$

Since $\frac{1}{|w|}$ is a decreasing function, the magnitude of $\widehat{\Delta L}(w)$ for higher frequency w is dampened compared to $|\widehat{\tau}(w)|$, which have been observed to be small [4]. The same analysis applies to linear momentum p since p only changes when there is a net external force $F \neq 0$ acting on the system, specifically $F = \frac{dp}{dt}$.

3.3 Total momentum loss

Generated motions must not only match real kinematics but also exhibit consistent dynamics. While motion models capture kinematic patterns, they often neglect dynamics such as ground reaction forces and torques, which are hard to measure at scale. To address this, we introduce a loss term that aligns the linear and angular momentum profiles of the generated motions with the ground truth. Additionally, we constrain the power spectrum of the whole-body angular momentum to match that of the real data, following the discussion in Section 3.2.

We define Δ_{AMo} and Δ_{LMo} as the difference between the momentum terms of recovered motion and ground truth:

$$\Delta_{\text{AMo}} = \text{AMo}(\hat{R}, \hat{\theta}) - \text{AMo}(R, \theta), \tag{3a}$$

$$\Delta_{\text{LMo}} = \text{LMo}(\hat{R}, \hat{T}, \hat{\theta}) - \text{LMo}(R, T, \theta), \tag{3b}$$

where $\text{AMo}(R, \theta)$ is defined in Eq. 2 and $\text{LMo}(R, T, \theta)$ is defined in Eq. 1. We define the components of our loss as followed:

$$\mathcal{L}_{\text{AMo}} = \|\Delta_{\text{AMo}}\|_2 + \|\Delta'_{\text{AMo}}\|_2, \tag{4a}$$

$$\mathcal{L}_{\text{LMo}} = \|\Delta_{\text{LMo}}\|_2 + \|\Delta'_{\text{LMo}}\|_2, \tag{4b}$$

$$\mathcal{L}_{\text{S}} = \|\mathcal{F}(\text{AMo}(\hat{R}, \hat{\theta})) - \mathcal{F}(\text{AMo}(R, \theta))\|_2, \tag{4c}$$

where f' denotes the time derivative of f and $\mathcal{F}(f)$ is f in frequency domain. Our proposed additional loss term for training motion models is

$$\mathcal{L}_{\text{TMo}} = \lambda_{\text{AMo}}\mathcal{L}_{\text{AMo}} + \lambda_{\text{LMo}}\mathcal{L}_{\text{LMo}} + \lambda_{\text{S}}\mathcal{L}_{\text{S}}. \tag{5}$$

We use the discrete Fourier transform $\mathcal{F}$ and the discrete cosine transform [51] for the spectrum loss $\mathcal{L}_S$.

4 Experiments

To evaluate the effectiveness of our loss function in increasing the consistency between global trajectory and local joint movement, we conduct experiments on the task of estimating global motion from the observations of joint motion in the root frame. This task is closely related to global motion reconstruction [50], a topic that has recently gained significant attention. In this setting, we exclude any information that could be inferred from the background to more clearly isolate the impact of our loss function.

We then demonstrate that incorporating our loss function into an existing method that leverages background information - such as camera movement - for global motion reconstruction improves motion plausibility.

Datasets. Following GLAMR [67], for the first task, we train baseline and our variants on AMASS [34]. We evaluate on EMDB [23] and the Kungfu subset of Motion-X [30]. For the second task, following WHAM [50], we pretrain the models on AMASS [34] and finetune on 3DPW [57], Human3.6M [19], MPI-INF-3DHP [37] and InstaVariety [21]. We evaluate on EMDB [23] and RICH [18]. For detailed datasets description, please refer to our supplementary.

Evaluation metrics. There are two different aspects to the performance of a motion recovery system: accuracy, reflected by root translation error (RTE); and plausibility, measured by foot skating, acceleration and jitter. For more detailed description and formulae, please refer to our supplementary material.

4.1 Global trajectory from local joint rotations

Baseline. In this task, we evaluate the performance of the global trajectory predictor from GLAMR [67] against a version of the same predictor trained with our loss function $\mathcal{L}_{\text{TMo}}$. We also compare the performance of our CVAE based predictor against PhysPT [72], which recovers plausible global motion from an initial kinematics estimate by training a transformer model with both kinematics-based and dynamics-based loss terms.

Quantitative results. Table 1 shows that our loss function improves over baselines for both plausibility linked metrics and global trajectory accuracy. While PhysPT [72] achieves comparable jitter, the deviation of its momentum profile from the ground truth shows that there are still implausible motions generated.

Table 1: The predicted root trajectory from ground truth local pose. PhysPT† denotes the method with the same Transformer architecture as PhysPT [72], using only position based loss and our global trajectory predictor.

Models	EMDB 2			Kungfu		
	RTE↓	Jitter↓	FS↓	RTE↓	Jitter↓	FS↓
GLAMR	5.19	13.70	7.49	9.89	40.56	6.86
GLAMR + $\mathcal{L}_{\text{TMo}}$	**4.73**	7.70	**4.91**	**9.16**	**34.20**	**4.81**
PhysPT	15.15	8.74	6.64	12.47	36.43	5.90
PhysPT†	7.43	8.36	11.15	23.82	38.06	10.66
PhysPT† + $\mathcal{L}_{\text{TMo}}$	6.58	**5.20**	8.22	13.62	35.06	7.17

Qualitative results. A typical improvement in body stability with our loss is shown in Figure 1 (rightmost), even though this metric was not specifically targeted. The improved stability comes from a lack of sudden large spikes in angular momentum, which is usually observed in sequences with this type of error. See Figure 2 (left) for the distribution change in the error of swing around gravity, indicative of body balance.

PhysPT [72] focuses on human-ground contact and neglects the conservation of momentum. This can be most clearly observed in sequences involving jumping, where the foot sliding and jitter measures cannot effectively identify impossible motions such as the bounce in mid-jump shown in Figure 2 (right).

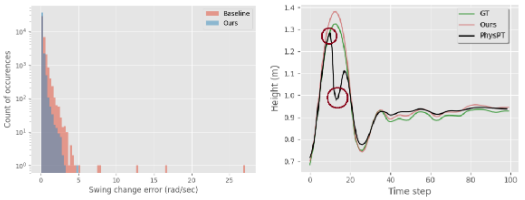

Figure 2: Left: improvement in body stability, reflected by error of the change in the swing around gravity. Right: Root joint height during a jumping sequence. PhysPT [72] predicts unnatural changes of the root trajectory along the gravity direction, highlighted by red circles.

Perceptual study. To assess the increase in plausibility due to our loss, we conduct a human study on Prolific [1]. We select 40 sequences from AMASS and EMDB where predictions made with our version of GLAMR differ the most from the baseline. Participants are shown several videos, each consisting of two animated motions displayed side by side. They are asked to choose the motion they find more plausible.

For each sequence, we collect 3 comparisons: ground truth vs. baseline, ours vs. baseline, and ground truth vs. ours. Each comparison is rated by 25 participants, with the same participants evaluating all comparisons for a given source sequence. The study includes two catch trials, where ground truth motions are compared with severely corrupted motions. No participants failed the catch trials.

There is a clear preference for our result compared to the baseline, with ours rated better for more than 70% of the time, while baseline is preferred in less than 15% of cases. Binomial test confirms that the preference for our method over the baseline is significant ($p < 0.05$). This is not solely due to the poor quality of the baseline predictions, however. We are rated as better or equal to the ground truth more than 60% of the time.

4.2 Global motion recovery

Baseline. In this task, we use WHAM [50] as the baseline.

Quantitative results. SOTA comparisons for this setting is shown in Tab. 2. Baseline WHAM [50] already outperforms optimization-based method SLAHMR [64] and regression-based TRACE [52], but with our loss, it was able to reduce jitter and foot sliding error. While TRAM [59] performs much better in global trajectory accuracy (RTE), its lack of consideration for the motion dynamic in global space leads to large errors in plausibility linked metrics.

Qualitative results. Figure 3 discusses some typical errors that our loss can address.

4.3 Ablation study

In this section, we analyze the effect of different loss functions on the performance of the global motion inference system GLAMR [67].

Table 2: SOTA comparisons for global human motion recovery. * denotes results taken from WHAM [50]. GLAMR† uses local pose results of WHAM.

Models	EMDB 2			RICH		
	RTE↓	Jitter↓	FS↓	RTE↓	Jitter↓	FS↓
GLAMR† [67]	6.5	19.9	5.7	8.8	17.9	4.0
TRACE* [52]	17.7	2987.6	370.7	610.4	1578.6	230.7
SLAHMR* [64]	10.2	31.3	14.5	28.9	34.3	5.1
TRAM [59]	**1.4**	109.1	20.3	-	-	-
WHAM [50]	4.1	21.0	4.4	**4.1**	19.7	3.3
GLAMR† [67] + $\mathcal{L}_{\text{TMo}}$	5.6	**15.1**	**3.6**	8.6	**13.4**	**2.4**
WHAM [50] + $\mathcal{L}_{\text{TMo}}$	4.3	19.7	3.7	4.4	17.9	3.2

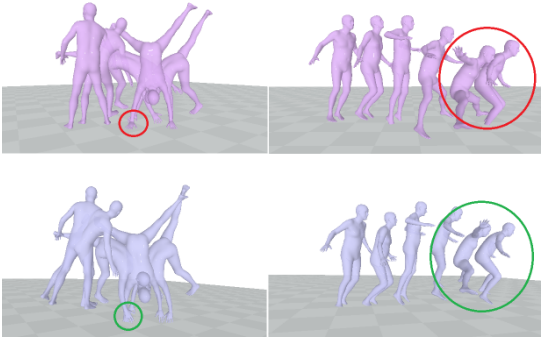

Figure 3: Qualitative examples of our improvement over WHAM [50]. Left: during cartwheel motion, baseline did not keep the hand in correct contact with ground. Right: During a jump, baseline did not reason correctly about the height of the body's CoM.

Components We investigate the effect of each component of $\mathcal{L}_{\text{TMo}}$, results in table 4. Observe that $\mathcal{L}_{\text{TMo}}$ performs well for all metrics, consistently ranked first compared to the rest. Regulating $\mathcal{L}_{\text{LMo}}$ is equivalent to the control of CoM trajectory, which helps reduce root trajectory errors. We compare this loss term to directly regulating joint velocity $\mathcal{L}_{\text{Jv}}$ (table 3). While both approaches show similar performance in terms of jitter and foot sliding, $\mathcal{L}_{\text{LMo}}$ outperforms $\mathcal{L}_{\text{Jv}}$ in reducing RTE. Inspection of root rotation swing error shows that while $\mathcal{L}_{\text{LMo}}$ can remove large outliers, $\mathcal{L}_{\text{AMo}}$ is more effective at reducing this error overall.

Other related terms Next, we compare our proposed loss to related terms to highlight its advantages. The results are reported in Table 3.

1. Root rotation can be decomposed into the twist around the gravity direction and the remaining swing. While twist can undergo drastic change quickly, swing is connected to the stability of the pose and tends to remain more stable. We propose the term $\mathcal{L}_{\text{SW}}$ which minimizes change between the swing rotations of consecutive frames. While reducing the sort of balance error observed in Figure 1, compared to baseline, this loss performs worse in all reported metrics, showing that our loss does more than simply smoothing out swing rotations.

2. Is it necessary to consider the shape of the body parts? To answer this question, we model the human body as a collection of point masses concentrated at the centroid of each body part and guide the model to match the angular momentum of such a system.

$$\text{TF}(R,\theta) = \sum_{i=1}^{P} m_i c_i^B \times \dot{c}_i^B \tag{6a}$$

$$\mathcal{L}_{\text{TF}} = \|\text{TF}(\hat{R},\hat{\theta}) - \text{TF}(R,\theta)\|_2 \tag{6b}$$

Compare the performance of this loss to $\mathcal{L}_{\text{AMo}}$, we observe that $\mathcal{L}_{\text{AMo}}$ outperforms it in all metrics, suggesting that the contribution of the angular momentum around the body parts' own CoM (and thus, the shape of the body parts) is indeed significant.

Table 3: Ablation study on other related losses.

	RTE↓	Jitter↓	FS↓
$\mathcal{L}_{\text{SW}}$	4.54	25.21	8.00
$\mathcal{L}_{\text{TF}}$	4.26	19.24	6.71
$\mathcal{L}_{\text{Jv}}$	4.60	19.31	6.55

Table 4: Ablation study on the loss components.

$\mathcal{L}_{\text{LMo}}$	$\mathcal{L}_{\text{AMo}}$	$\mathcal{L}_{\text{S}}$	RTE↓	Jitter↓	FS↓
			4.19	23.33	7.50
✓			3.90	19.06	6.51
	✓		4.35	15.52	5.11
		✓	4.23	15.53	5.19
✓	✓		3.85	13.40	4.10
✓	✓	✓	3.82	12.64	3.65

4.4 Analysis

Low data regime. We investigate the effect of the amount of data available for training on the performance. We train GLAMR trajectory predictor with and without our loss on a random sample of the AMASS training set, with 20%, 50% and 70% of the size of the total set. To facilitate comparisons between different variants, we introduce a composite measure, m_{AB} defined between two variants A and B as followed

$$m_{\text{AB}} = \frac{\text{RTE}_A}{\text{RTE}_B} \times \frac{\text{Jitter}_A}{\text{Jitter}_B} \times \frac{\text{FS}_A}{\text{FS}_B} \tag{7}$$

Setting the performance of the baseline at full size as the reference, from Table 5 we can see that with our loss the method can reach relatively good performance on fewer data.

Table 5: Effect of amount of available data on performance.

Training data percentage	0.2	0.5	0.7	1
GLAMR	3.21	1.49	1.39	1
GLAMR + $\mathcal{L}_{\text{TMo}}$	0.47	0.28	0.27	0.25
Gap	6.87	5.37	5.22	4.03

Table 6: The amount of extra time required by our loss compared to baseline.

Baseline	Percentage	Absolute
GLAMR [67]	30.1%	6h
WHAM [50]	11.0%	2h
TEMOS [41]	2.2%	$<$ 1h
PhysPT† [72]	3.2%	$<$ 1h

Training time. One drawback of our loss is increased time during training. The size of the effect varies by baseline, specifically whether the components we need for calculation of the momentum terms are available, or need to be computed on the fly. The relatively large increase observed in GLAMR's trajectory predictor [67] is also due to the relatively simple loss terms used by the baseline, which only involves the root trajectory.

5 Conclusion

We have introduced TMo, a simple yet effective loss function based on momentum features of human motion, designed to improve the physical plausibility of reconstructed motion. By encouraging consistency between global body movement and local joint dynamics through linear and angular momentum alignment, TMo addresses common artifacts such as foot sliding and unrealistic body acceleration. Applied to the global motion recovery task, TMo consistently enhances motion quality across multiple models, yielding results that are more physically realistic than baseline approaches—without compromising accuracy or requiring architectural changes.

Limitations and future directions While our loss function is versatile, its application requires retraining existing models, which can be time-consuming and reliant on access to external codebases. Future work could focus on developing optimization methods that leverage a self-supervised version of our loss for easier integration. Another promising avenue is incorporating environmental context into the momentum profile optimization process.

References

[1] URL https://www.prolific.com/.

[2] Fabien Baradel, Romain Br'egier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez. PoseBERT: A Generic Transformer Module for Temporal 3D Human Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12798–12815, 2022.

[3] Emad Barsoum, John R. Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–149909, 2017.

[4] Wendy L. Boehm, Kieran M Nichols, and Kreg G. Gruben. Frequency-dependent contributions of sagittal-plane foot force to upright human standing. *Journal of biomechanics*, 83:305–309, 2019.

[5] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your Commands via Motion Diffusion in Latent Space. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2022.

[6] Enrico Chiovetto, Meghan E. Huber, Dagmar Sternad, and Martin A. Giese. Low-dimensional organization of angular momentum during walking on a narrow beam. *Scientific Reports*, 8, 2018.

[7] Lin Cong, Philipp Ruppel, Yizhou Wang, Xiang Pan, Norman Hendrich, and Jianwei Zhang. Efficient Human Motion Reconstruction from Monocular Videos with Physical Consistency Loss. *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[8] Erik Gartner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable Dynamics for Articulated 3d Human Motion Reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13180–13190, 2022.

[9] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory Optimization for Physics-Based Reconstruction of 3d Human Pose from Monocular Video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13096–13105, 2022.

[10] D. Geman and S. Geman. Bayesian Image Analysis. In E. Bienenstock, F. F. Soulié, and G. Weisbuch, editors, *Disordered Systems and Biological Organization*, NATO ASI Series, vol. 20, pages 709–743. Springer, Berlin, Heidelberg, 1986. doi: 10.1007/978-3-642-82657-3_30.

[11] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and Tracking Humans with Transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14737–14748, 2023.

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating Diverse and Natural 3D Human Motions from Text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022.

[13] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. NeMF: Neural Motion Fields for Kinematic Animation. In *NeurIPS*, 2022.

[14] Dorian F. Henning, Tristan Laidlow, and Stefan Leutenegger. BodySLAM: Joint Camera Localisation, Mapping, and Human Motion Tracking. *ArXiv*, abs/2205.02301, 2022.

[15] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow. *ACM Transactions on Graphics (TOG)*, 39:1 – 14, 2019.

[16] Hugh Herr and Marko Popovic. Angular momentum in human walking. *Journal of Experimental Biology*, 211:467 – 481, 2008.

[17] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural MoCon: Neural Motion Control for Physically Plausible Human Motion Capture. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6407–6416, 2022.

[18] Chun-Hao Paul Huang, Hongwei Yi, Markus Hoschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and Inferring Dense Full-Body Human-Scene Contact. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13264–13275, 2022.

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

[20] Michael Kallay. Computing the Moment of Inertia of a Solid Defined by a Triangle Mesh. *Journal of Graphics Tools*, 11:51 – 57, 2006.

[21] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D Human Dynamics From Video. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, 2018.

[22] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing Diffusion Noise Can Serve As Universal Motion Priors. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1334–1345, 2023.

[23] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14586–14597, 2023.

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2019.

[25] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and Camera Motion Estimation from in-the-wild Videos. *2024 International Conference on 3D Vision (3DV)*, pages 397–408, 2023.

[26] Tae-Joung Kwon, Yoonsang Lee, and Michiel van de Panne. Fast and flexible multi-legged locomotion using learned centroidal dynamics. *ACM Transactions on Graphics (TOG)*, 39:46:1 – 46:17, 2020.

[27] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-Generic Hierarchical Human Motion Prior using VAEs. *2021 International Conference on 3D Vision (3DV)*, pages 771–781, 2021.

[28] 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.

[29] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *European Conference on Computer Vision*, 2022.

[30] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. *Advances in Neural Information Processing Systems*, 2023.

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2015.

[32] Sihan Ma, Qiong Cao, Hongwei Yi, Jing Zhang, and Dacheng Tao. GraMMaR: Ground-aware Motion Model for 3D Human Motion Reconstruction. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[33] Adriano Macchietto, Victor B. Zordan, and Christian R. Shelton. Momentum control for balance. *ACM SIGGRAPH 2009 papers*, 2009.

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture As Surface Shapes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019.

[35] Galo Maldonado, François Bailly, Philippe Souéres, and Bruno Watier. On the coordination of highly dynamic human movements: an extension of the uncontrolled manifold approach applied to precision jump in parkour. *Scientific Reports*, 8, 2018.

[36] Ángel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose Transformers (POTR): Human Motion Prediction with Non-Autoregressive Transformers. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2276–2284, 2021.

[37] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2016.

[38] Zaka-Ud-Din Muhammad, Zhangjin Huang, and Rashid Khan. A review of 3D human body pose estimation and mesh recovery. *Digit. Signal Process.*, 128:103628, 2022.

[39] Takuo Negishi and Naomichi Ogihara. Regulation of whole-body angular momentum during human walking. *Scientific Reports*, 13, 2023.

[40] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021.

[41] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

[42] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT Motion-Language Dataset. *Big Data*, 4(4):236–252, dec 2016. doi: 10.1089/big.2016.0028.

[43] Marko B. Popovic, Andreas G. Hofmann, and Hugh M. Herr. Angular momentum regulation during human walking: biomechanics and control. *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, 3:2405–2411 Vol.3, 2004.

[44] Darcy S. Reisman, John P. Scholz, and Gregor Schöner. Coordination underlying the control of whole body momentum during sit-to-stand. *Gait & posture*, 15 1:45–55, 2002.

[45] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan C. Russell, Ruben Villegas, and Jimei Yang. Contact and Human Dynamics from Monocular Video. In *Symposium on Computer Animation*, 2020.

[46] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D Human Motion Model for Robust Pose Estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11468–11479, 2021.

[47] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human Motion Prediction via Spatio-Temporal Inpainting. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7133–7142, 2018.

[48] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap. *ACM Transactions on Graphics (TOG)*, 39:1 – 16, 2020.

[49] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick P'erez, and Christian Theobalt. Neural monocular 3D human motion capture with physical awareness. *ACM Transactions on Graphics (TOG)*, 40:1 – 15, 2021.

[50] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing World-Grounded Humans with Accurate 3D Motion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2023.

[51] Gilbert Strang. The Discrete Cosine Transform. *SIAM Rev.*, 41:135–147, 1999.

[52] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8856–8866, 2023.

[53] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model. *ArXiv*, abs/2209.14916, 2022.

[54] Shashank Tripathi, Lea Muller, Chun-Hao Paul Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D Human Pose Estimation via Intuitive Physics. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023.

[55] Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael J. Black, Daniel Holden, and Carsten Stoll. HUMOS: Human Motion Model Conditioned on Body Shape. In *European Conference on Computer Vision (ECCV)*, 2024.

[56] Jaap H. van Dieën, Sjoerd M. Bruijn, Koen K. Lemaire, and Dinant A. Kistemaker. Simultaneous stabilizing feedback control of linear and angular momentum in human walking. *bioRxiv*, 2025.

[57] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision*, 2018.

[58] Miomir Vukobratovic and Branislav Borovac. Zero-Moment Point - Thirty Five Years of its Life. *Int. J. Humanoid Robotics*, 1:157–173, 2004.

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

[60] Keenon Werling, Janelle Kaneda, Alan Tan, Rishi Agarwal, Six Skov, Tom Van Wouwe, Scott Uhlrich, Nicholas Bianco, Carmichael F. Ong, Antoine Falisse, Shardul Sapkota, Aidan Chandra, Joshua Autton Carter, Ezio Preatoni, Benjamin Fregly, Jennifer Hicks, Scott L. Delp, and C. Karen Liu. AddBiomechanics Dataset: Capturing the Physics of Human Motion at Scale. In *European Conference on Computer Vision*, page 490–508, 2024.

[61] Alexander W. Winkler, C. Dario Bellicoso, Marco Hutter, and Jonas Buchli. Gait and trajectory optimization for legged systems through phase-based end-effector parameterization. *IEEE Robotics and Automation Letters*, 3:1560–1567, 2018.

[62] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based Human Motion Estimation and Synthesis from Videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11512–11521, 2021.

[63] Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Yu Wang, Xinchao Wang, and Yanfeng Wang. EqMotion: Equivariant Multi-Agent Motion Prediction with Invariant Interaction Reasoning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1420, 2023.

[64] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling Human and Camera Motion from Videos in the Wild. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21222–21232, 2023.

[65] Ye Yuan and Kris Kitani. Residual Force Control for Agile Human Behavior Imitation and Extended Motion Synthesis. *ArXiv*, abs/2006.07364, 2020.

[66] Ye Yuan and Kris M. Kitani. DLow: Diversifying Latent Flows for Diverse Human Motion Prediction. In *European Conference on Computer Vision*, 2020.

[67] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11028–11039, 2021.

[68] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason M. Saragih. SimPoE: Simulated Character Control for 3D Human Pose Estimation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7155–7165, 2021.

[69] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-Guided Human Motion Diffusion Model. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15964–15975, 2022.

[70] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[71] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning Motion Priors for 4D Human Body Capture in 3D Scenes. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11323–11333, 2021.

[72] Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. PhysPT: Physics-aware Pretrained Transformer for Estimating Human Dynamics from Monocular Videos. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2305–2317, 2024.

[73] Yufei Zhang, Jeffrey O. Kephart, and Qiang Ji. Incorporating Physics Principles for Precise Human Motion Prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6164–6174, January 2024.