GraMMaR: Ground-aware Motion Model for 3D Human Motion Reconstruction
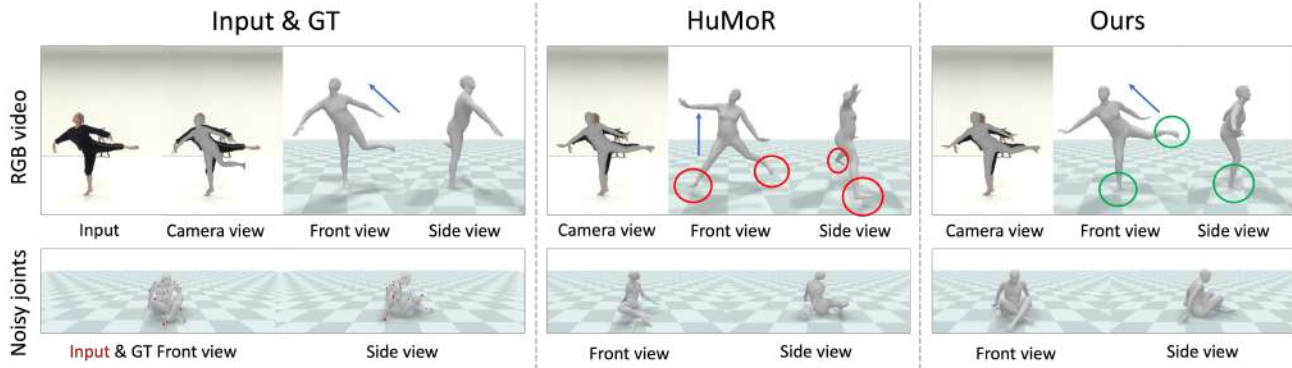
Anonymous Author(s)

Submission Id: 2526

Figure 1: 3D motion in the camera view is misleading. A representative optimization method HuMoR [34] produces correct poses under camera view but physically implausible poses in world view when faced with ambiguity (Row1) and noise (Row2). In contrast, our method provides a realistic motion while being ground-aware. To achieve it, we explicitly present the human-ground interaction and devise a CVAE-based model to jointly learn the distribution of transitions in interaction and pose. Body torso direction and contacts for HuMoR and ours are highlighted. GT in Row1 is reconstructed from multi-view images.

ABSTRACT

Demystifying complex human-ground interactions is essential for accurate and realistic 3D human motion reconstruction from RGB videos, as it ensures consistency between the humans and the ground plane. Prior methods have modeled human-ground interactions either implicitly or in a sparse and limited manner, often resulting in unrealistic and incorrect motions when faced with noise and uncertainty. In contrast, our approach explicitly represents these interactions in a dense and continuous manner. To this end, we propose a novel **Gr**ound-**a**ware **M**otion **M**odel for 3D Human Motion **R**econstruction, named **GraMMaR**, which jointly learns the distribution of transitions in both pose and interaction between every joint and ground plane at each time step of a motion sequence. It is trained to explicitly promote consistency between the motion and distance change towards the ground. After training, we establish a joint optimization strategy that utilizes GraMMaR as a dual-prior, regularizing the optimization towards the space of plausible ground-aware motions. This leads to realistic and coherent motion reconstruction, irrespective of the assumed or learned ground plane. Through extensive evaluation on the AMASS and AIST++ datasets, our model demonstrates good generalization and discriminating abilities in challenging cases including complex and ambiguous human-ground interactions. The source code will be released.

CCS CONCEPTS

• **Computing methodologies → Motion capture**.

KEYWORDS

Motion reconstruction, 3D human motion

1 INTRODUCTION

The human body frequently engages in movements that involve interactions with the ground plane. In real-life scenarios, when a body part is in close proximity to the ground, individuals may need to slow down, lean their torso, orient their head to look at the ground or position their hands and feet on the ground. The capability to accurately and plausibly predict 3D human motion from RGB videos, which encompasses realistic interactions with the ground, is crucial for numerous applications [45], such as scene understanding [8, 19, 43], 3D dance motion reconstruction and generation [5, 20, 21, 39], and augmented and virtual reality

games [4, 32, 33]. While extensive research has focused on 3D motion estimation under camera space[1, 41], considering alignment solely in the camera view might be insufficient and potentially deceptive. There are cases where poses appear reasonable in the camera view but exhibit physically implausible body support on the assumed ground plane when viewed from an alternate viewpoint or placed in a 3D scene. Moreover, even within the camera view, handling noisy observations can result in visually implausible recovered motions, such as body twists, penetration, and jittery movements. Fig. 1 demonstrates these issues. These challenges primarily arise because most state-of-the-art methods rarely consider the interaction between humans and the ground plane.

To address these issues, a natural solution is to model human-ground interaction explicitly to ensure consistency between the human body and the ground. Essentially, human-ground interaction involves the interdependent relationships between a 3D human and the ground plane. However, to date, few methods have explored human-ground interaction; those that do have primarily focused on the body-ground contact using binary contact labels [34, 35] or ground reaction force [41]. [35] models the interaction by directly predicting binary contact labels, indicating whether predefined joints are in contact with the ground. These classification results are then used as hard constraints that restrict the distance between joints and the ground during inference, thereby enabling the generation of physically plausible poses. However, the use of binary labels is inadequate as it only applies to joints in direct contact with the ground, leaving most other joints without physical restrictions. Moreover, the sparse and uncertain nature of contact occurrence across all joints significantly impacts the accuracy of motion reconstruction. The performance heavily relies on the quantity of frames and joints within a given motion sequence where contact is established, making it unstable. Alternatively, some work [41] has introduced ground reaction force as a means of representing human-ground interaction. A larger reaction force corresponds to a heavier penalty on the distance between the joints and the ground during optimization. Although intuitive, it is difficult to access and only applies to joints in contact with the ground.

In this work, we address these issues by building a robust human motion model that accurately captures the dynamics of 3D human motion through human-ground interactions. To achieve this goal, we first introduce a novel continuous distance-based per-joint interaction representation to encode fine-grained human-ground interactions at each time step. It overcomes the limitations of binary contact labels and ground reaction force by combining per-joint ground distance and its velocity along the gravity axis. Unlike previous methods, our new representation provides a continuous and differentiable measure with physical significance, allowing for a comprehensive depiction of motion patterns and ground-based body support for both contacting and non-contacting joints.

Building upon the novel representation, we devise an explicit ground-aware motion dynamics model that incorporates human-ground interactions and human pose. This is formulated as an autoregressive conditional variational auto-encoder (CVAE) [38] to capture the temporal variations in human pose and human-ground interaction. The model simultaneously learns the distribution of transitions for both pose and joint-to-ground distances across adjacent frames within a motion sequence, producing a wide range

of plausible poses and human-ground interactions. By conditioning the decoder to predict future motion based on existing poses and human-ground interactions, the model enforces consistency between the body and the ground plane.

We train our model on AMASS [26] and develop a joint optimization strategy for 3D human motion reconstruction from noisy observations and RGB videos. The trained model serves as a dual-prior to regularize the optimization towards the space of plausible ground-aware motions, resulting in realistic and coherent motion reconstruction and generation, regardless of the assumed or learned ground plane. The resultant reconstruction method is termed GraM-MaR, which stands for **Gr**ound-**a**ware **M**otion **M**odel for 3D Human Motion **R**econstruction. We evaluate GraMMaR quantitatively and qualitatively on both RGB videos and noisy settings and demonstrate its superiority over the baseline in complex and ambiguous contact conditions. GraMMaR proves effective irrespective of the ground plane being known or unknown.

2 RELATED WORK

Kinematic estimation. Kinematic methods for 3D pose estimation in videos [2, 3, 6, 9, 10, 16, 17, 23, 24, 28, 29, 34, 42, 46] can be categorized as end-to-end learning-based or optimization-based approaches. End-to-end methods, such as VNect [29], directly extract 2D and 3D joint positions using CNN-based regression, while VIBE [17] estimates SMPL body model [24] parameters using a temporal generation model trained with a discriminator. Other works, such as LEMO [46] and HuMoR [34], train priors for motion transition and human-scene relations using large-scale motion data [13, 15, 26, 30], which can be used for fitting 3D poses from observations during optimization. However, these methods may produce physically implausible results, such as body twists and foot skating, especially for complex actions or when training data is limited.

Physics-based estimation with simulators. Several methods [11, 12, 14, 25, 44] have been proposed to enhance physical plausibility by incorporating physics laws. These methods use physics simulators such as MuJoCo [40] and Isaac Gym [27] as a black box to guide 3D pose prediction. Due to the non-differentiable nature of the physics simulator, reinforcement learning is employed to learn control of the simulator [25, 44]. For instance, SimPoE [44] uses a kinematic-aware policy to generate control signals for the physics simulator to recover realistic 3D poses. Similarly, PoseTriplet [12] incorporates the simulator into a semi-supervised framework to reduce artifacts in pseudo labels. Although effective, they can be computationally intensive for training from scratch and prone to collapse, limiting their generalization ability on videos in the wild. To address this issue, differentiable simulators such as TDS [11] are introduced for articulated 3D motion reconstruction.

Physics-based estimation without simulators. Recent research [7, 35, 37, 41, 46] has focused on developing physical constraints for 3D motion optimization that do not require physics simulators [7, 35, 37, 41, 46]. These methods learn to predict contact conditions for specific joints, imposing boundary constraints during optimizationGraviCap [7] incorporates the physical properties of moving objects in a scene to recover scale, bone length, and ground simultaneously. However, these constraints are only applied to contact joints and overlook the physical characteristics
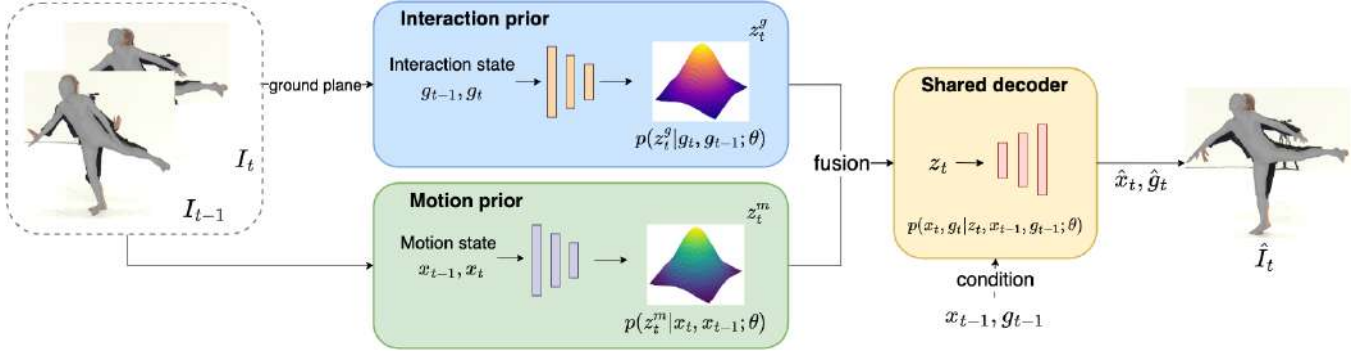
Figure 2: *GraMMaR architecture.* In training, given the previous state I_{t-1} and current state I_t, we obtain the motion state x_{t-1}, x_t, and interaction state g_{t-1}, g_t. Our model learns the transition of motion and interaction state changes separately by two priors and reconstructs $\hat{x}_t$, $\hat{g}_t$ by sampling from the two distributions and decoding them conditioned on both x_{t-1} and g_{t-1}.
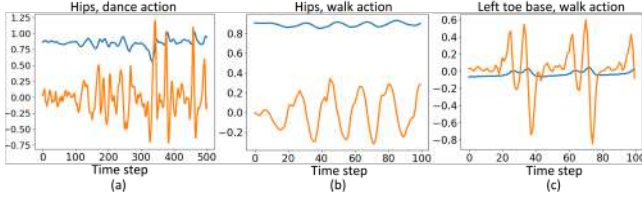

Figure 3: t-SNE analysis of the innteraction state g in Eq. 1. We see its components d and v present unique and dense patterns in separating different types of motion in (a)-(b) and different joints in the same motion sequence in (b)-(c).

of the body's other joints. [41] infers reaction forces from contact joints and forwards them to the entire body via dynamic equations, but this approach results in approximation errors. In our work, we propose a continuous representation of human-ground interaction that enables us to investigate interaction conditions for all joints, including non-contact ones.

Human-ground interaction representation in pose estimation. In physics-based methods for pose estimation [34–36, 46], human-ground interaction is typically defined as foot-ground contact to impose constraints on height, velocity, and ground reaction forces during optimization. Alternatively, [41] introduced a soft contact variable related to penetration distances and contact forces, to penalize penetrations and excessive accelerations on 3D joints. However, these methods only consider binary contact and ignore non-contact joints. To address this limitation, we propose a continuous and expressive representation for human-ground interaction and establish CVAE-based generative model for human-ground relations to achieve physically plausible motions and reasonable ground planes.

3 METHOD

We propose GraMMaR, a robust generative motion model that captures the dynamics of 3D human motion while being ground-aware,

and demonstrate its effectiveness as a regularizer in optimization-based approaches for estimating accurate and plausible 3D human motion and ground plane.

Preliminaries. We represent the state of a person by an interaction state g defined in the subsequent section, and a motion state x following [34]. The motion state x is composed of a root translation $r \in \mathbb{R}^3$, a root orientation $\Phi \in \mathbb{R}^3$, body pose joint angles $\Theta \in \mathbb{R}^{3 \times 21}$, joint positions $J \in \mathbb{R}^{3 \times 22}$ and their velocities. All the angles are in axis-angle format.

3.1 Human-ground Interaction Analysis

Representation for human-ground interaction state. In contrast to binary contact labels, our objective is to devise a more comprehensive and detailed representation that indicates not only whether a joint contacts the ground, but also characterizes the interaction state between joints and the ground in the present, past, and, most importantly, immediate future. This will enable capturing information regarding joints approaching the ground, moving away from the ground, and remaining stationary in the air.

To this end, we represent the human-ground relationship as g, consisting of the human-to-ground distance $d \in \mathbb{R}^{23}$ between all joints and the ground including the root joint, as well as its velocity $v \in \mathbb{R}^{23}$ along the gravity axis. We employ SMPL body model [24] and utilize only the first 23 joints for calculations, *i.e.*,

$$g = [d, v]. \tag{1}$$

To calculate the human-to-ground distance, we use either the assumed ground or the ground variable $n \in \mathbb{R}^4$ to be optimized at each step, which will be discussed in Section 3.3. With the ground plane n, we can get a random point Q on it. For the i-th joint J_i, there is an angle α_i between the ground plane normal $\vec{n}_d \in \mathbb{R}^3$ and vector $\overrightarrow{QJ}_i$. By calculating the projection of vector $\overrightarrow{QJ}_i$ on the plane normal $\vec{n}_d$, we can get the distance representation as follows,

$$d^i = |\overrightarrow{QJ}_i| \cdot cos(\alpha_i), \ d = [d^0, d^1, \ldots, d^{22}]. \tag{2}$$

Moreover, we assume that the ground is flat, rigid, and has a floor normal vector oriented along the gravitational axis. In this work, we also make the assumption that the human body primarily
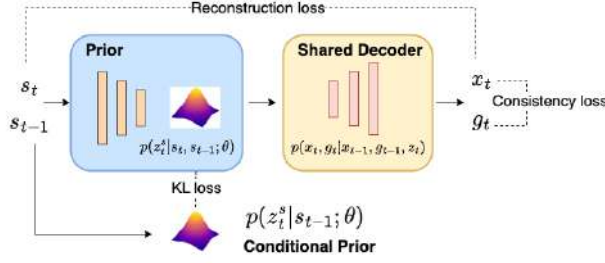
Figure 4: Training of GraMMaR. For simplicity, "Prior" can be either interaction prior or motion prior. Similarly, s_t can indicate g_t and x_t, depending on the prior type.

interacts with the ground, a circumstance encountered in most in-the-wild cases, such as dance, yoga, and other activities.

Analysis of the interaction state. As shown in Fig. 3, compared with general motion dynamic modeling, our interaction state g, including distance d and distance velocity v, presents unique and dense patterns in separating different types of motion (Fig. 3(a)-(b)) and different joints in the same motion sequence (Fig. 3(b)-(c)).

Comparison with binary contact label. Compared to the binary contact label, the continuous interaction representation as shown in Fig. 3 provides more detailed information beyond mere contact. For example, suppose the distance between a joint and the ground is zero, and the velocity along the gravitational direction is significant. It indicates that the joint has recently made contact with the ground, and due to inertia, both the joint and its adjacent counterparts are likely to continue moving toward the ground for the next few frames. Under these conditions, it is highly improbable for the joints to exhibit any motion in the opposite direction.

3.2 Ground-aware Generative Motion Model

Building upon the proposed representation, we aim to develop an explicit ground-aware motion dynamics model that incorporates human-ground interactions with human pose to capture the temporal variations in human pose and human-ground interaction.

Specifically, we model the probability of a motion sequence x_t by considering the human-ground interaction g_t at each step, *i.e.*,

$$p_\theta(x_0, g_0, x_1, g_1, \cdots, x_T, g_T) = p_\theta(x_0, g_0) \prod_{t=1}^{T} p_\theta(x_t, g_t | x_{t-1}, g_{t-1}), \quad (3)$$

where x_t and g_t are the motion and interaction states at the time step, respectively. For each time step, the overall motion depends not only on the motion state x_{t-1} at the previous time step but also on the human-ground interaction state formulated as g_{t-1}. Consequently, this allows $p_\theta(x_t, g_t | x_{t-1}, g_{t-1})$ to capture the fine-grained physical plausibility of the transition.

As illustrated in Fig. 2, we propose GraMMaR that leverages a conditional variational autoencoder (CVAE) to represent the transition probability. This model formulates the probability of transition in motion state and interaction state as follows:

$$p_\theta(x_t, g_t | x_{t-1}, g_{t-1}) = P_\theta(z_t | x_{t-1}, g_{t-1}) \cdot p_\theta(x_t, g_t | x_t), \quad (4)$$

where z_t is the latent variable for time step t. For the purpose of computation efficiency, we formulate $p_\theta(z_t | x_{t-1}, g_{t-1})$ into two independent encoders as:

$$
\begin{aligned}
p_\theta(z_t^m | x_{t-1}), p_\theta(z_t^g | g_{t-1}), \\
z_t = z_t^m \oplus z_t^g,
\end{aligned}
\quad (5)
$$

where z_t^m and z_t^g denote the latent variables for motion and human-ground interaction respectively, and $\oplus$ denotes the concatenation operation. These two encoders are approximated as independent Gaussian distributions using implicit neural networks. To enable the differentiation and learning of unique characteristics for the two encoders, we employ two conditional priors as guidance rather than relying on the standard Gaussian distribution, *i.e.*,

$$
\begin{aligned}
p_\theta(z_t^m | x_{t-1}) = \mathcal{N}(z_t^m; \mu_\theta(x_{t-1}), \sigma_\theta(x_{t-1})), \\
p_\theta(z_t^g | g_{t-1}) = \mathcal{N}(z_t^g; \mu_\theta(g_{t-1}), \sigma_\theta(g_{t-1})).
\end{aligned}
\quad (6)
$$

By simultaneously learning the distribution of transitions for both pose and joint-to-ground interactions across adjacent frames within a motion sequence, our model can produce a wide range of plausible poses while being ground-aware.

In the next step, we employ **a shared decoder** to estimate the future motion conditioned on both the motion state and the human-ground interaction from the previous step, thereby ensuring consistency between the body pose and the ground plane. Specifically, the shared decoder is designed to enable the combination of multiple inputs, including a random motion latent sample z_t^m, a random interaction latent sample z_t^g (with the combined latent variables denoted as z_t), the motion state x_{t-1}, and the interaction representation g_{t-1}. Besides, to facilitate an auto-regressive manner in further applications, it outputs the motion state x_t and the interaction g_t simultaneously. Compared with the baseline, it also predicts a binary contact label c_t for the predefined nine contact joints.

Training loss and implementation details. The training loss contains reconstruction loss $\mathcal{L}_{recon}$ for motion state and interaction state, KL loss $\mathcal{L}_{KL}$ between conditional prior and the corresponding encoder output, and the consistency loss $\mathcal{L}_{consist}$ between motion state and learned interaction state, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{recon} + \mathcal{L}_{KL} + \mathcal{L}_{consist}, \quad (7)$$

where the reconstruction loss $\mathcal{L}_{recon}$ is defined as:

$$
\begin{aligned}
L_{recon} = L_{recon\_m} + L_{recon\_g}, \\
L_{recon\_m} = ||x_t - \hat{x}_t||^2, L_{recon\_g} = ||g_t - \hat{g}_t||^2,
\end{aligned}
\quad (8)
$$

given the training pair $(x_t, g_t, x_{t-1}, g_{t-1})$. $\hat{x}_t$ and $\hat{g}_t$ are the output of the decoder for motion state and interaction state, respectively. The KL loss $\mathcal{L}_{KL}$ is calculated separately for motion and interaction states by computing the KL divergence $D_{KL}(\cdot||\cdot)$ between the output of the encoder and the corresponding conditional prior. The consistency loss $L_{consist}$ promotes consistency between the learned interaction state $\hat{g}_t$ and the human-ground interaction information, which is extracted through the predicted joints $\hat{x}_t$ and the learned ground plane $\hat{n}$, *i.e.*,

$$L_{consist} = ||\hat{g}_t - f(\hat{x}_t, \hat{n})||^2. \quad (9)$$

Lastly, for comparing the contact accuracy with the baseline, we also incorporate a contact classification head and compute the BCE loss between the predicted contact label $\hat{c}_t$ and the ground truth.

3.3 Joint Optimization Strategy

After training, following [34], we devise a joint optimization strategy for 3D human motion reconstruction from noisy observations and RGB videos. We leverage GraMMaR to regularize optimization toward the space of plausible ground-aware motions, thereby maintaining consistency between the human body and the ground plane. We consider our GraMMaR for two tasks: (1) denoising under the fixed ground plane; and (2) motion reconstruction from RGB videos where the ground plane is unavailable and subject to optimization alongside the motion sequence.

Optimization variables. Given a sequence of motion observations $y_{0:T}$ in 2D/3D joints format and an optional ground plane n, we aim to obtain a sequence of SMPL parameters $(r_{0:T}, \Phi_{0:T}, \Theta_{0:T})$, body shape β, and ground plane n (if not provided), which could not only match the observation, but also maintain physical plausibility and consistency between human and ground. Our GraMMaR could be incorporated into the optimization by parameterizing the SMPL parameter sequence into an initial motion state x_0, an initial interaction state g_0, and a sequence of latent variables $z_{0:T}$ composed of motion latent variables $z_{0:T}^m$ and interaction latent variables $z_{0:T}^g$. With optimized latent variables and initial states, we can roll-out the whole sequence of SMPL parameters through the decoder in an auto-regressive way. Please see supplementary for more details.

Noisy observation setting. In this setting, we consider a scenario where a ground plane and a set of noisy joint positions, generated using existing motion reconstruction algorithms like SMPLify [1], are available in a noisy form. Our goal is to optimize the motion sequence to ensure both physical plausibility and accuracy in human-ground interactions when the ground plane is provided and fixed. We show our model performs better when used for fitting to noisy joints and known ground plane, especially for the challenging cases.

To this end, the objective function is formulated as a combination of dual-prior loss, prior consistency loss, data loss, and regularization loss, with the last two loss terms following the design of [34]. In this context, we primarily focus on the dual-prior loss:

$$L_{prior} = L_{prior\_m} + L_{prior\_g},$$

$$L_{prior\_m} = \prod_{t=1}^{T} log\mathcal{N}(z_t^m; \mu_\theta(x_{t-1}), \sigma_\theta(x_{t-1})),$$

$$L_{prior\_g} = \prod_{t=1}^{T} log\mathcal{N}(z_t^g; \mu_\theta(g_{t-1}), \sigma_\theta(g_{t-1})),$$

$$(10)$$

and the prior consistency loss:

$$L_{pconsist} = \hat{g}_t - f(\hat{x}_t, \hat{n}), \qquad (11)$$

where $\hat{g}_t$ and $\hat{x}_t$ are the output of the shared decoder with the input latent variables z_t^g and z_t^m. $\hat{n}$ is the fixed ground plane normal.

RGB video setting. In this particular setting, we tackle the problem of motion reconstruction from RGB videos, where a set of 2D/3D keypoint positions $y_{0:T}$, extracted from each individual frame in the camera view, is provided, but without any knowledge of the ground plane. Our objective is to seek both the physically plausible and precise motion state $x_{0:T}$ and the ground plane n that can transform the motion state into world space.

In contrast to the noisy observation setting, the ground plane is unavailable this time. As a result, we establish a ground plane variable n, allowing it to be optimized alongside the motion state. In total, we optimize the motion and interaction latent variables $z_{0:T}^m$, $z_{0:T}^g$, initial motion and intearction states x_0, g_0, and the ground plane vector n at same time. In each optimization iteration, the prior consistency loss, as shown in Eq. (11), is calculated based on the optimized ground plane vector $\hat{n}$ rather than the assumed one.

Implementation details. During optimization, initialization phase includes three parts, latent variables $(z_{1:T}^m, z_{1:T}^g)$, first-frame latent variables z_0^m, z_0^g, and first-frame motion and interaction states x_0, g_0. We first initialize the SMPL parameters for the whole sequence by an single-frame algorithm, and thus obtain the initialization of first frame states x_0, g_0 and the latent variables $(z_{1:T}^m, z_{1:T}^g)$ through the trained encoders $p_\theta(z_t^m|x_{t-1})$, $p_\theta(z_t^g|g_{t-1})$. For the first frame latent variables z_0^m, z_0^g, we adopt a GMM distribution following [34] to initialize z_0^m, and use a random initialization for z_0^g. Our model shows that random initialization is sufficient for z_t^g.

4 EXPERIMENT

4.1 Datasets and Splits

AMASS [26] is a large motion capture dataset containing multiple types of motions, mainly running, walking, and turning around. We follow [34] to process the sequences into 30 hz and extract the contact labels for evaluation. Our model and the baseline are both trained on the training set of AMASS and evaluated on the test set of all datasets without retraining or fine-tuning.

To assess the effectiveness of our proposed model in handling various types of human-ground relationships, we partition the test set of AMASS into distinct levels according to the minimum hip height within each sequence. Our hypothesis is that as the minimum hip height decreases, the interaction between the human and the ground becomes increasingly intricate.

AIST++ dataset [22] comprises a vast collection of dance motion data that includes RGB videos, multiple camera parameters, and 3D motion annotations for 1,408 sequences of 30 different subjects. For the purpose of evaluating our model's performance under different human-ground relations, we partition the test set according to the degree of difficulty involved in estimating the ground plane. Further information regarding this partitioning is discussed in the subsequent section and supplementary material.

4.2 Baselines and Metrics

Baselines. We conducted a comparative analysis between our method and the baseline HuMoR [34], a CVAE-based prior that does not take dense human-ground interaction into account. We ensure that the initialization and optimization settings are identical for both methods. In the noisy observation setting, VPoser-t serves as the initialization algorithm, while in the RGB video setting, we use PARE [18], a single-frame learning-based pose reconstruction technique, for initialization. VPoser-t uses VPoser [31] and 3D joints smoothness constraints during optimization.

Metrics. In our evaluation, we employ several common metrics to assess the performance of our method and the baseline. The 3D positional errors are measured by the mean per joint position error

Method	MPJPE-G ($\downarrow$)	MPJPE ($\downarrow$)	MPJPE-PA ($\downarrow$)	PVE ($\downarrow$)	contact acc ($\uparrow$)	accel mag ($\downarrow$)
VPoser-t [31]	32.8	34.8	27.9	43.2	-	61.6
VPoser-t + HuMoR [34]	22.7	23.9	19.0	30.3	89.3%	**16.7**
VPoser-t + **GraMMaR**	**21.9**	**22.6**	**18.1**	**29.5**	**91.1%**	20.8

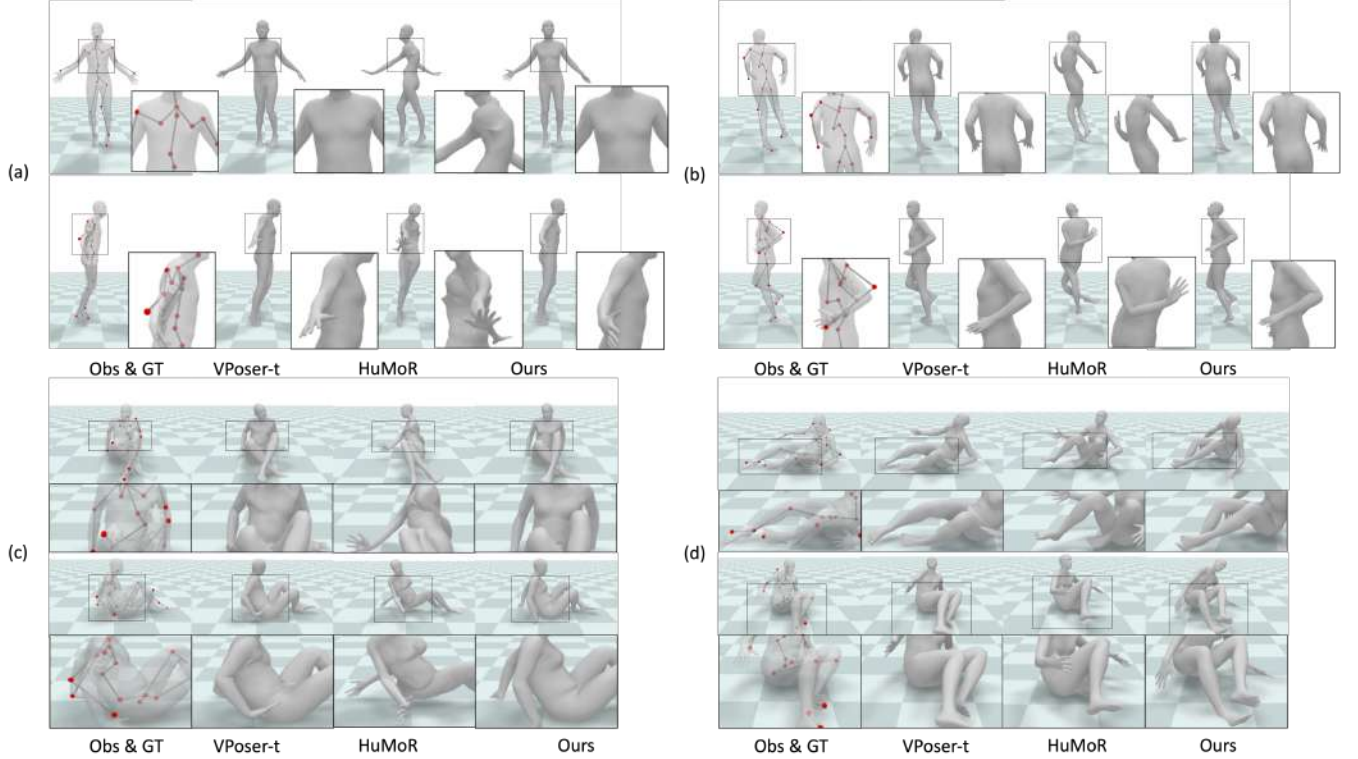**Table 1: Results on the AMASS dataset under the noisy observation setting.**

Figure 5: Qualitative comparison on the AMASS test set under the noisy observation setting. Our method doesn't show body twist even under complex human-ground interaction. For each case, the first row shows the front view in the world space, while the second row shows the side view. Please view the supplementary video and materials for more details.

(MPJPE), MPJPE after procrustes alignment (MPJPE-PA), MPJPE over global positions (MPJPE-G), and the per-vertex error (PVE). In addition, we evaluate the binary classification of nine pre-defined joints [34] that are likely to be in direct contact with the ground. We also assess the smoothness of the generated motion by computing the average per-joint accelerations (Accel). Moreover, we report the performance of our method on different levels of human-ground interaction, which cannot be captured by the overall errors on the entire test set. We also report the cosine similarity scores (Cos) to evaluate performance in estimating the ground plane.

4.3 Optimization with noisy observations

First, we evaluate GraMMaR with the observation of noisy 3D joint positions and a fixed ground plane, and demonstrate that GraMMaR performs better than the baseline, especially in cases with complex human-ground relations. We use the 90-frame (3s) clips from the AMASS dataset. To simulate the presence of noise, we introduce

Gaussian noise to the joint positions with a mean of zero and a standard deviation of 0.04m, following [34]. The categorization of different levels of human-ground interaction is based on the minimum height of the hip within the clip.

Table 1 presents the mean results attained over the entire test set of the AMASS dataset. We compare GraMMaR with baseline HuMoR, as well as the initialization method VPoser-t. Our results demonstrate that our GraMMaR approach produces more precise poses and yields better performance in terms of contact accuracy. These findings suggest that the use of interaction states facilitates the extraction of human-ground interaction and significantly enhances human-ground relations. Regarding smoothness, while HuMoR reports the lowest acceleration, our approach outperforms VPoser-t substantially and provides an inherently smooth outcome. In contrast to HuMoR, our method affords greater flexibility to accommodate noisy poses, particularly those characterized by complex human-ground relations.

Method	Metric	0-0.3	0.3-0.6	0.6-1.0	avg
VPoser-t [31]	MPJPE-G (↓)	**34.4**	33.5	32.6	33.5
	MPJPE (↓)	**38.0**	36.0	34.5	36.2
	PVE (↓)	**48.1**	44.4	43.0	45.2
VPoser-t + HuMoR [34]	MPJPE-G (↓)	53.0	24.0	21.6	32.9
	MPJPE (↓)	53.0	25.8	22.8	33.9
	PVE (↓)	65.3	33.1	28.9	42.4
	contact acc (↑)	75.9%	86.6%	90.0%	84.2%
VPoser-t + GraM-MaR	MPJPE-G (↓)	48.2	**22.6**	**21.1**	**30.6**
	MPJPE (↓)	49.1	**23.9**	**21.7**	**31.6**
	PVE (↓)	59.6	**30.4**	**28.5**	**39.5**
	contact acc (↑)	**78.6%**	**87.6%**	**91.8%**	**86.0%**

Table 2: Results on AMASS dataset under the noisy observation setting at different levels of human-ground interaction.
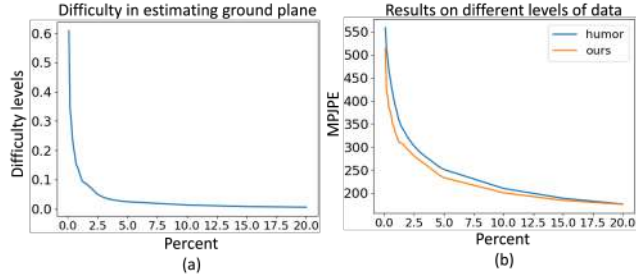

Figure 6: (a) Difficulty levels of the top 20% of challenging data. (b) Our method outperforms the baseline HuMoR in the top 20% of challenging cases.

Table 2 presents the outcomes for data splits categorized by varying levels of human-ground interaction. Our approach demonstrates superior performance, particularly in the most challenging level "0-0.3". At this level, our method exhibits improvements in both positional error and contact accuracy, indicating that it produces a more physically realistic and accurate pose with a more reasonable contact condition. While VPoser-t displays a consistently robust performance across all levels of data, it is unable to predict the ground plane and exhibits inferior smoothness capabilities.

Fig. 5 presents qualitative examples of our approach compared to HuMoR. In Figures 5(a) and 5(b), HuMoR exhibits body twist in jumping, while our method doesn't. As HuMoR lacks an understanding of human-ground interaction, it struggles to accurately discern the motion and distinguish between joint position changes caused by noise versus those caused by the actual action itself. In Figures 5(c) and 5(d), HuMoR shows inaccurate orientation and body twist in sitting because of complexity in motion and the human-ground interaction, while our method performs well in these cases.

4.4 Optimization with RGB video

Next, we show that our GraMMaR can predict a more physically reasonable pose and ground plane simultaneously, and can accurately figure out the ambiguous pose under camera view. In this setting, we use 60-frame (2s) video clips from the AIST++ dataset.

Besides the positional metrics and smoothness metrics, we use the cosine similarity scores (Cos) between normal vectors of planes to evaluate the estimation of the ground plane.

To quantify the challenge of estimating the ground plane for video clips, we assume that a larger divergence in the predicted ground plane by different methods indicates a higher level of difficulty in pose ambiguity. This is due to the fact that the pose in camera view may not readily differentiate the ground plane. To assess this, we calculate the cosine similarity scores of the predicted ground plane from our approach and the baseline HuMoR separately, and then sort clips according to the absolute difference in similarity scores between the two methods. The absolute difference for the top 20% of clips is shown in Fig. 6(a), while the remaining clips indicate a negligible difference and are therefore not presented.

Table 3 presents the results of our method, baseline HuMoR, and the initialization method PARE, for the entire test set and the top 1% of clips. Our approach exhibits superior performance in estimating the ground plane, particularly for the top 1% of difficult data regarding human-ground relations. This suggests our method can better distinguish between mistaken poses under camera view. In terms of smoothness, both HuMoR and our method show significant improvements compared to PARE, albeit at an acceptable cost of position accuracy. Although our approach reports relatively inferior results regarding positional metrics for the entire test set, it can produce more reasonable poses under the world space by considering the ground plane, especially for ambiguous motions. As shown in Fig. 6(b), regarding the MPJPE in world space, our GraMMaR outperforms HuMoR on the top 20% difficult cases.

The qualitative examples also provide evidence to support this conclusion. Fig. 7 presents some examples from AIST++. We showcase both the prediction in camera view and in world view Since PARE cannot predict the ground plane, we exclude its prediction under the world view. As demonstrated in Fig. 7, HuMoR generates accurate poses in most cases under camera view, but produces completely physically unreasonable poses with incorrect contact conditions in world space. This suggests that HuMoR is incapable of resolving ambiguous poses in world space and solely optimizes motion by observation. On the other hand, our method, aided by the interaction map, accurately resolves ambiguous poses and generates physically plausible poses with the correct conditions.

Generalizing to videos in the wild. Finally, we compare our method with the baseline HuMoR on videos sourced from the Internet and demonstrate that our method generalizes better to videos in the wild without the need for retraining or fine-tuning. Fig. 8 showcases that for challenging scenarios like yoga, dance, and handstanding, our method generates physically plausible poses with more reasonable human-ground relations.

5 LIMITATION AND FUTURE WORK

Although our model can yield superior performance in predicting physically plausible motion and reasonable ground planes in challenging cases, there are some limitations, such as its inability to handle hand penetration and cases involving a second contact plane. In some extreme cases, our method can make a reasonable inference on the ground plane but have a large error in positions due to the extreme angle and the high moving speed. Nonetheless,

Method	Cos (↑)	Cos 1% (↑)	Accel (↓)	Accel align (↓)	MPJPE-G (↓)	MPJPE (↓)	MPJPE-PA (↓)	MPJPE* 1% (↓)
PARE [18]	-	-	65.6	23.8	**257.3**	**102.5**	**62.0**	-
PARE + HuMoR [34]	0.99175	0.70452	**4.0**	**3.3**	606.2	114.3	80.7	383.0
PARE + **GraMMaR**	**0.99965**	**99956**	4.4	3.6	666.0	130.5	92.9	**327.0**

Table 3: Results on AIST++ dataset under the RGB video setting. "Cos" is the mean cosine similarity between the predicted ground plane and the gt. "Cos 1%" is the Cos scores of the top 1% difficult clips in estimating the ground plane. "MPJPE* 1%" denotes the MPJPE of the predictions in world space for the top 1% difficult data in estimating the ground plane.
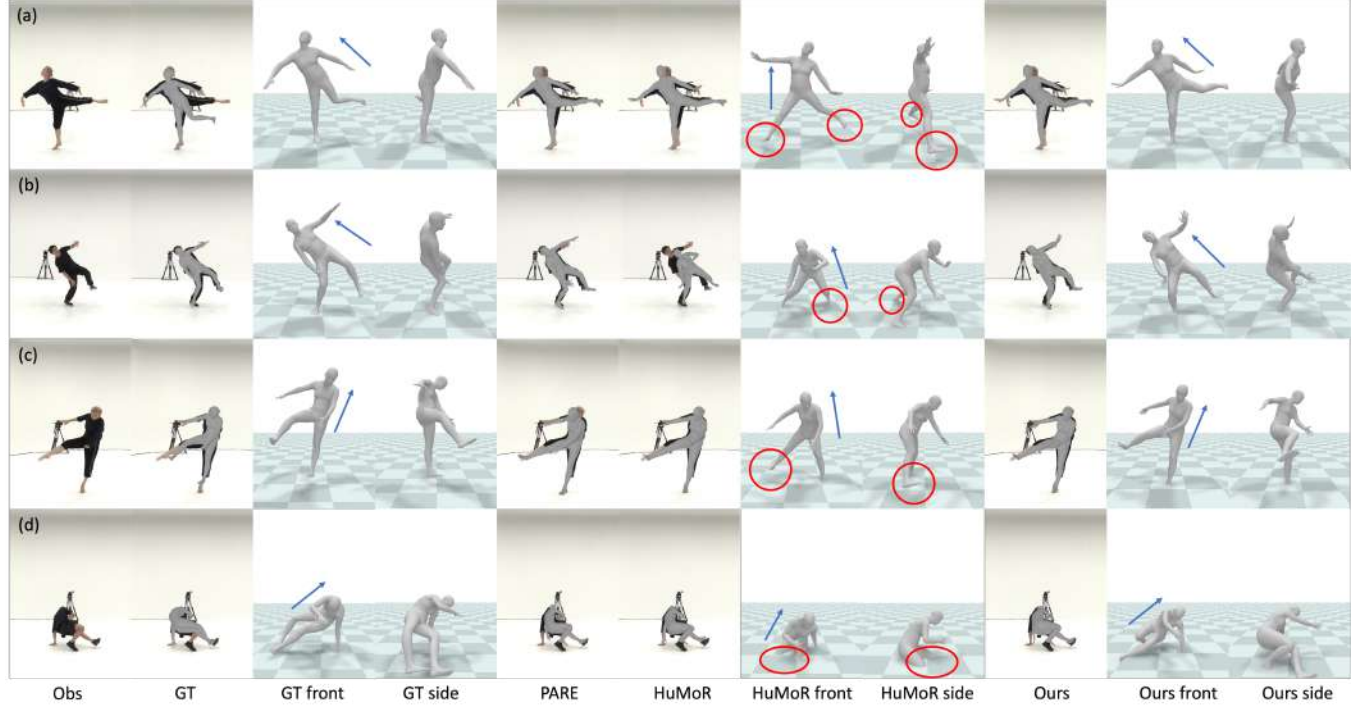

Figure 7: Qualitative comparison on the AIST++ test set under the RGB video setting. "front" and "side" denote the front and side view in world space. The direction of the body torso and contacts of HuMoR are highlighted. HuMoR tends to predict the body torso in a direction perpendicular to the ground while our method doesn't. Please see supplementary for more details.

our approach outperforms the baseline method HuMoR in these challenging cases (See supplementary for more details). In future work, it is promising to learn a stronger prior from large-scale training data (*e.g.*, containing multiple contact planes, fine-grained hand actions) to further improve the performance.

6 CONCLUSION

In this work, we propose a dense and continuous representation for human-ground interaction and a CVAE-based model named GraMMaR based on it to address the consistency issue between the human and the ground. We further establish a joint optimization-based approach that uses our proposed GraMMaR as a regularizer to estimate physically plausible and correct motion from noisy observations and RGB videos. The proposed method demonstrates promising results in generating realistic and physically plausible outcomes, particularly in challenging scenarios characterized by complex and ambiguous human-ground interaction.
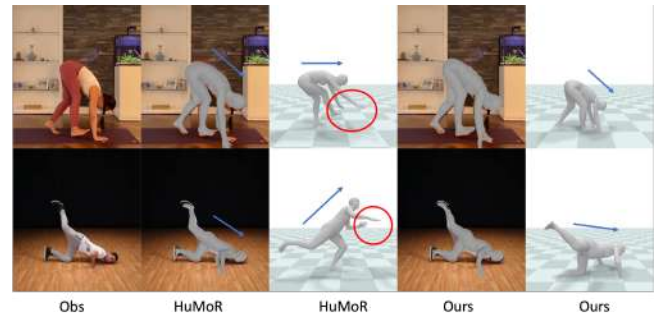

Figure 8: Qualitative comparison on the videos from the Internet. The blue arrow and red circle have the same meaning as in the above figures.

REFERENCES

[1] 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: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 561–578.

[2] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 2013. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3618–3625.

[3] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2272–2281.

[4] Polona Caserman, Augusto Garcia-Agundez, and Stefan Göbel. 2019. A survey of full-body motion reconstruction in immersive virtual reality applications. *IEEE transactions on visualization and computer graphics* 26, 10 (2019), 3089–3108.

[5] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. 2021. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

[6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. 2018. Learning 3d human pose from structure and motion. In *Proceedings of the European conference on computer vision (ECCV)*. 668–683.

[7] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. 2021. Gravity-aware monocular 3d human-object reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12365–12374.

[8] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. 2017. Blitznet: A real-time deep network for scene understanding. In *Proceedings of the IEEE international conference on computer vision*. 4154–4162.

[9] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. 2015. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3810–3818.

[10] Ahmed Elhayek, Carsten Stoll, Kwang In Kim, and Christian Theobalt. 2015. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 86–98.

[11] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. 2022. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13190–13200.

[12] Kehong Gong, Bingbing Li, Jianfeng Zhang, Tao Wang, Jing Huang, Michael Bi Mi, Jiashi Feng, and Xinchao Wang. 2022. PoseTriplet: co-evolving 3D human pose estimation, imitation, and hallucination under self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11017–11027.

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. 2019. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints. In *International Conference on Computer Vision*. 2282–2292. https://prox.is.tue.mpg.de

[14] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. 2022. Neural MoCon: Neural Motion Control for Physically Plausible Human Motion Capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6417–6426.

[15] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. 2022. Capturing and Inferring Dense Full-Body Human-Scene Contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 13274–13285.

[16] Sena Kiciroglu, Helge Rhodin, Sudipta N Sinha, Mathieu Salzmann, and Pascal Fua. 2020. Activemocap: Optimized viewpoint selection for active human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 103–112.

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5253–5263.

[18] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. 2021. PARE: Part Attention Regressor for 3D Human Body Estimation. In *Proceedings International Conference on Computer Vision (ICCV)*. IEEE, 11127–11137.

[19] Taein Kwon, Bugra Tekin, Siyu Tang, and Marc Pollefeys. 2022. Context-Aware Sequence Alignment using 4D Skeletal Augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8172–8182.

[20] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1272–1279.

[21] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

[22] 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]

[23] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. 2020. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5064–5073.

[24] 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.

[25] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. 2021. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems* 34 (2021), 25019–25032.

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

[27] 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.

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *Acm Transactions On Graphics (TOG)* 39, 4 (2020), 82–1.

[29] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)* 36, 4 (2017), 1–14.

[30] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J. Mitra. 2019. iMapper: Interaction-guided Scene Mapping from Monocular Videos. *ACM SIGGRAPH* (2019).

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

[32] Ken Pfeuffer, Matthias J Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. 2019. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[33] Francesco Pilati, Maurizio Faccio, Mauro Gamberi, and Alberto Regattieri. 2020. Learning manual assembly through real-time motion capture for operator training with augmented reality. *Procedia Manufacturing* 45 (2020), 189–195.

[34] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11488–11499.

[35] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. 2020. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 71–87.

[36] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. 2021. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–15.

[37] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. 2020. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)* 39, 6 (2020), 1–16.

[38] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[39] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. 2020. DeepDance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia* 23 (2020), 497–509.

[40] 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

[41] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. 2021. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11532–11541.

[42] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. 2020. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on computer vision and Pattern recognition*. 899–908.

[43] Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. 2022. Human-aware object placement for visual environment reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3959–3970.

[44] 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*. 7159–7169.

[45] Jing Zhang and Dacheng Tao. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal* 8, 10 (2020), 7789–7817.

[46] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. 2021. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11343–11353.