HumanMM: Global Human Motion Recovery from Multi-shot Videos

Anonymous CVPR submission

Paper ID 2019

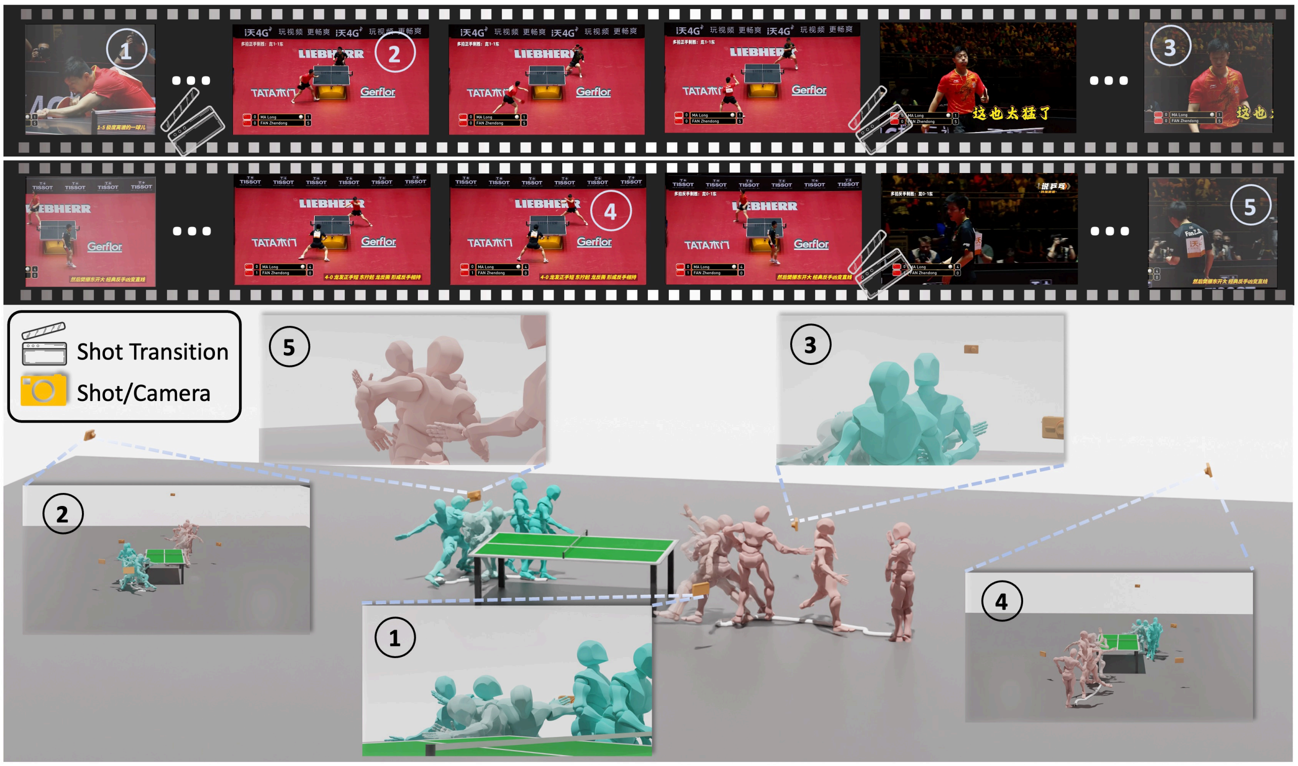

Figure 1. **Recovering a human motion from multi-shot videos.** **Top**: We take two multi-shot table tennis game videos with shot transitions as input. We aim to recover two motions of two athletes (Long MA and Zhendong FAN) from two videos, respectively. The first video is recorded by three shots ("①", "②", and "③"), and the second one is recovered by two shots ("④" and "⑤"). **Bottom**: We recover two motions (Long MA in green and Zhendong FAN in pink), different shots, and camera poses for each multi-shot video. The recovered motion is aligned with the motion in the videos.

Abstract

In this paper, we present a novel framework designed to reconstruct long-sequence 3D human motion in the world coordinates from in-the-wild videos with multiple shot transitions. Such long-sequence in-the-wild motions are highly valuable to applications such as motion generation and motion understanding, but are of great challenge to be recovered due to abrupt shot transitions, partial occlusions, and dynamic backgrounds presented in such videos. Existing methods primarily focus on single-shot videos, where continuity is maintained within a single camera view, or simplify multi-shot alignment in camera space only. In this work, we tackle the challenges by integrating an enhanced camera pose estimation with Human Motion Recovery (HMR) by incorporating a shot transition detector and a robust alignment module for accurate pose and orientation continuity across shots. By leveraging a custom motion integrator, we effectively mitigate the problem of foot sliding and ensure temporal consistency in human pose. Extensive evaluations on our created multi-shot dataset from public 3D human datasets demonstrate the robustness of our method in reconstructing realistic human motion in world coordinates.

1. Introduction

In recent years, significant advances have been made in 3D human pose estimation, particularly in enhancing the accuracy of human motion recovery (HMR)[1] from monoc-

[1] In this paper, the "human mesh recovery" refers to recovery in the camera coordinates and the "human motion recovery" denotes recovery in the world coordinates. Unless specified otherwise, HMR refers to **human motion recovery**.

ular video sequences. HMR has demonstrated extensive applications in areas such as human-AI interaction [1, 2], human motion understanding [3–6], and motion generation [3, 4, 7–25]. While existing methods [26, 27] have achieved relatively high performance in recovering mesh in camera coordinates, estimating human motion in world coordinates remains challenging [28–31] due to inaccurate camera pose estimation and the complexity of reconstructing human motion spatially.

Most current progress in 3D human motion community mainly benefits from large scale data [26, 27, 29–33], and long-sequence videos. These resources enhance estimation accuracy for HMR methods and improve the understanding and generation of longer motion sequences for tasks such as motion understanding [3, 34, 35] and generation [3, 4, 7–25, 35–51], even when annotations are derived from markerless capturing methods like pseudo labels [52–55].

A promising approach to enlarge the scale of the motion databases is to estimate human motions from *unlimited* online videos in a *markerless* manner. However, many long-sequence online videos are recorded with multiple shots, referred to as multi-shot videos[2], especially prevalent in domains such as sports broadcasting, talk shows, and concerts. In filmmaking and television live show, a “shot” denotes an individual camera view capturing a specific moment or action from a particular vantage point [56].

Segmenting multi-shot videos into separate shots inevitably reduces the length of the video sequences, which can be detrimental to tasks that benefit from longer sequences, such as long motion generation [51, 57]. This limitation is highlighted in the existing datasets [58, 59], where the longest clip is less than 20 seconds after segmentation, as shown in Fig. 2. Moreover, focusing exclusively on online single-shot videos diminishes the utilization ratio of available online videos and may negatively impact the diversity of scenarios represented in the created datasets.

Therefore, *how to address the issue of discontinuities caused by shot transitions* is notoriously difficult in the community. To resolve this problem, previous works [60–63] have proposed algorithms to address human mesh recovery in a camera space from movies containing shot change between long shots and close-ups.

However, recovering human motions in world coordinates from multi-shot videos presents two fundamental challenges that remain underexplored. 1) *How to align the human motion and orientation in the world coordinates during shot transitions?* Ensuring continuity of human orientation and pose across shots is complicated by factors such as partial visibility of human body (*e.g.* transitioning from long shot to close-up) and changes in human orientation (*e.g.* two long shots from different viewpoints). These issues, caused by abrupt changes in camera viewpoints, necessitate robust alignment mechanisms. 2) *How to reconstruct accurate human motion in world coordinates?* Existing approaches employ Simultaneous Localization and Mapping (SLAM) methods to estimate camera parameters, which are then used to project recovered human meshes from camera to world coordinates [28–31]. This process requires highly accurate camera estimation and must address motion consistency and foot sliding in the recovered human motion within the world space.

[2]In this paper, a **multi-shot video** refers to a long-sequence video containing multiple shot transitions. We assume that the camera intrinsics remain consistent across different shots within a multi-shot video.

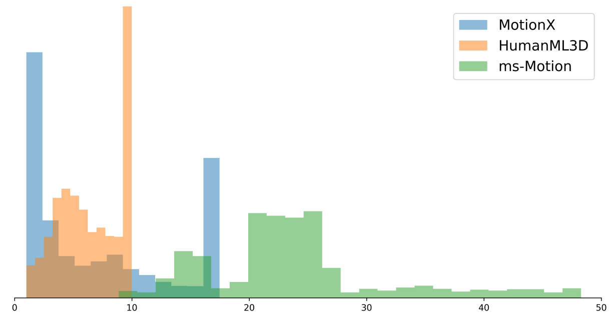

Figure 2. The comparison between the distribution of sequence lengths in different existing large-scale markerless motion datasets with ours. The x-axis and y-axis denote the duration time (s) and percentage of video number, respectively. Our dataset (in green) contains more portion of long-sequence videos in general.

Despite these challenges, human motion in multi-shot videos often remain continuous across shots, even as camera viewpoints change. This observation suggests that with appropriate handling of shot transitions and camera motion, it is possible to reconstruct consistent and complete 3D human motions throughout multi-shot videos.

In this paper, we propose a novel framework *HumanMM*, Human Motion recovery from Multi-shot videos, to address these challenges. It integrates human pose estimation across shots with robust camera estimation in the world space. First, we develop a shot transition detector to identify frames with shot transitions. To ensure a more robust camera pose estimation, we introduce an enhanced SLAM method incorporating long-term tracking of feature points and exclusion of moving human from bundle adjustment process. We utilize existing HMR method integrated with our enhanced camera estimation to get the initial human parameters for each separated shot. Subsequently, we implement an alignment module to align human orientation based on stereo calibration and smooth human poses through a trained multi-shot HMR encoder, which effectively captures the temporal context of human movements across different shots. Finally, after aligning human and camera parameters between shot transitions, we train a motion decoder and a trajectory refiner to smooth the human pose and mitigate issues such as foot sliding, thereby enhancing the overall motion consistency in the reconstructed 3D human motions.

Our contributions can be summarized as follows.

- We present the first approach to reconstruct human motion from multi-shot videos in world coordinates.
- We introduce *HumanMM*, a HMR framework for multi-shot videos. It includes an enhanced camera trajectory estimation method, a human motion alignment module and a motion integrator to ensure accurate and consistent recovery of human pose and orientation in world coordinates across different shots in the whole video.
- We develop a multi-shot video dataset *ms*-Motion to evaluate the performance of HMR from multi-shot videos, based on existing public datasets such as AIST [64] and Human3.6M [65]. Extensive experiments on related benchmarks verify the effectiveness of our method.

2. Related Work

2.1. HMR from One-shot Video

One-shot videos, captured with a single camera without shot transitions, has been extensively studied within the community for human mesh and motion recovery.

Human mesh recovery in camera coordinates can be broadly categorized into two approaches: optimization-based methods [66–70] and regression-based methods [32, 71–74]. With the significant advancements of transformer [75], HMR2.0 [26] has surpassed previous methods and benefits several downstream tasks related to HMR.

Although there are several previous works tried to recover motions in world coordinates with multi-camera capture system [64, 76] and IMU-based methods [77, 78] and enjoy relatively satisfying results, this setup limits their use for applications of *infinite* in-the-wild monocular videos. To address this limitation, several attempts [28–31] integrate SLAM into the HMR pipeline by first estimating the camera pose using SLAM methods, *e.g*. DROID-SLAM [79] or DPVO [80], and then project the recovered human motion from camera to world coordinates. To exclude the inconsistencies caused by dynamic objects, such as moving humans, TRAM [29] modifies DROID-SLAM by incorporating human masking and depth-based distance rescaling. However, DROID-SLAM performs dense bundle adjustment (DBA) on feature maps from downsampled images and selects features based only on two consecutive frames rather than long-term video sequences [79–81]. Consequently, masking significantly reduces the number of informative and consistent features, especially when humans occupy large portions of the image, leading to inaccuracies. Therefore, developing a SLAM method that retains sufficient and representative features for DBA after masking is important.

2.2. HMR from Multi-shot Video

Multiple shots are fundamental elements of cinematic storytelling and live performances, utilizing various camera positions and focal lengths to create immersive and detailed viewing experiences for audiences [56]. However, most marker-based motion capture (MoCap) datasets [64, 76, 77, 82, 83] consist single-shot videos only, resulting in limited research on HMR from multi-shot videos.

Recovering human motion from multi-shot videos in camera coordinates is already challenging. This is because treating each pose estimation result of each shot separately leads to inconsistencies when combining all estimations, caused by partially or fully invisible human bodies across shot transitions. Pavlakos *et al*. [60] addresses this issue by focusing on shot changes from long shots to close-ups, which are common in film. They develop smoothness constraints within a temporal Human Mesh and Motion Recovery (t-HMMR) model to infer motions during occlusions caused by shot transitions. Advancements in HMR methods [31] for single-shot videos in world coordinates have paved the way for extending HMR to multi-shot videos with varying camera viewpoints. However, aligning human orientation, body pose, and translation continuously across multi-shot videos in world coordinates underexplored. Effective alignment is crucial to maintain motion continuity and coherence, especially when dealing with diverse camera perspectives and abrupt transitions between shots.

In summary, while substantial progress has been made in HMR from single-shot videos, extending these techniques to multi-shot videos requires addressing additional complexities related to camera pose alignment and motion consistency across shot transitions. We address this challenge by proposing a novel pipeline that ensures accurate and continuous 3D HMR from multi-shot monocular videos.

3. Method

In this section, we propose *HumanMM* to recover human motion from multi-shot videos. The system overview is shown in Fig. 3. Given an input video sequence $\mathbf{V} = \{I_t\}_{t=1}^{T}$ of length T, where I_t denotes the t-th frame, our objective is to recover human motion in world coordinates. We begin by detecting shot transition frames based on human bounding box (*a.k.a*. bbox) and 2D keypoints (*a.k.a*. KPTs) through a *shot transition detector* (Sec. 3.2). For each clipped shot, we initialize the camera pose (camera rotation and camera translation) and recover initial human motion in world coordinates (Sec. 3.3). The initialized SMPL parameters and camera poses are then fed into a *human motion alignment* module (Sec. 3.4), which aligns human orientations via camera calibration based on human 2D KPTs and smooth the human pose by incorporating pose information across different shots. Additionally, it refines the entire motion sequence through whole video using a temporal motion encoder *ms*-HMR. Finally, we introduce a post-processing module for motion integration (Sec. 3.5).

3.1. Preliminary: 3D Human Model

Our method aims to recover motions in world coordinates in the SMPL [86] format, whose pose at frame t can be

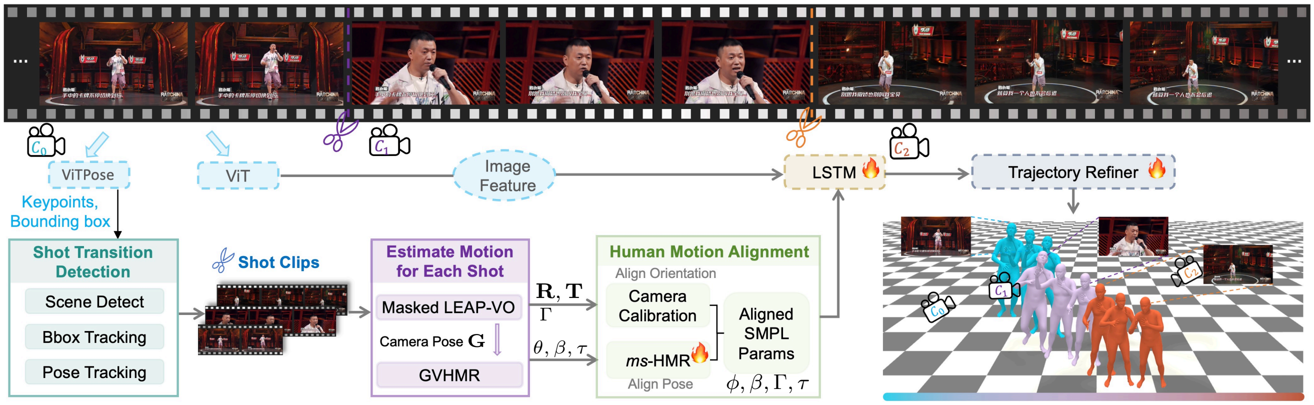

Figure 3. **The overview of *HumanMM*.** *HumanMM* processes multi-shot video sequences by first extracting motion feature such as keypoints and bounding boxes, using ViTPose [84] and image feature using ViT [85]. These features are then segmented into single-shot clips via *Shot Transition Detection* (Sec. 3.2). Initialized camera (camera rotation $\mathbf{R}$ and camera translation $\mathbf{T}$) and human (SMPL) parameters for each shot are estimated using *Masked LEAP-VO* (Sec. 3.3) and GVHMR [31]. Human orientation is aligned across shots through *camera calibration* (3.4.1), and *ms-HMR* (Sec. 3.4.2) ensures consistent pose alignment. Finally, a bi-directional *LSTM-based motion decoder* with *trajectory refiner* enhances motion consistency and mitigates foot sliding throughout the video.

represented as $\mathcal{M}_t(\theta_t, \beta_t, \Gamma_t, \tau_t) \in \mathbb{R}^{6890\times3}$. Here, the body pose, body shape, root orientation, and translation are $\theta_t \in \mathbb{R}^{23\times3}$, $\beta_t \in \mathbb{R}^{10}$, $\Gamma_t \in \mathbb{R}^3$, and $\tau_t \in \mathbb{R}^3$, respectively. We use $\mathbf{K}_t^{2D}$ to denote human 2D KPTs at each frame t.

3.2. Shot Transition Detector For Multi-shot Video

Our algorithm begins with shot transition detection in one video. As shown in Fig. 3, the *shot transition detector* has three key components, scene transition detector, bounding box (*a.k.a.* bbox) tracking, and human keypoints tracking. (1) *Scene change transition detector.* Initially, we employ the SceneDetect [87] algorithm to identify scene changes based on significant variations in the background. However, the SceneDetect fails to detect shot transitions when background changes are unnoticeable, illustrated in Fig. 4. Subsequently, we leverage the following modules to bridge the gap. (2) *Bbox tracking for shot transition.* As a shot change often accompanies with a sudden change of human subject size, we track humans in a video via mmtracking [88]. Consequently, we compute the Intersection over Union (IoU) between neighbor bboxes and identify a shot transition when the IoU falls smaller than a manually tuned threshold. (3) *Human pose tracking for shot transition detection.* To achieve a finer granularity, we additionally introduce human 2D KPTs to detect extreme corner shot changes in a video. By thresholding the IoU of corresponding keypoints between neighbor frames, we can accurately identify shot transitions even with subtle human movements.

As each separate module cannot identify all kinds of shot transitions, the three modules are jointly used to clip a video into several sub-sequences serially.

3.3. Human Motion and Camera Pose Estimation For Each Shot

After obtaining the clipped videos, our next goal is to estimate the camera pose and SMPL parameters in the world

(a) Scene Change (b) Position Change (c) Pose Change

Figure 4. **Shot transition detection examples**. Examples (a), (b), and (c) illustrate multi-shot scenarios in online videos. (a) shows scene transitions detectable by SceneDetect. (b) illustrates significant position changes undetectable by SceneDetect but resolvable with bbox tracking-based method. (c) shows pose or orientation transition, requiring pose tracking-based methods as they cannot be addressed by either SceneDetect or bbox tracking.

coordinates for each clipped video. The estimated camera pose and motions for each shot will be used to construct the whole motion sequence in the next stage (Sec. 3.4).

How to estimate the camera parameters accurately? Our approach for camera parameter calculation is based on a visual odometry (VO) estimation method, LEAP-VO [81]. Utilizing the CoTracker method [89], LEAP-VO estimates the visibility and trajectories of N selected points by analyzing image gradients across the video sequence. LEAP-VO subsequently computes confidence scores for each trajectory, retaining only those with high confidence while discarding trajectories shorter than a predefined threshold. The remaining trajectories undergo bundle adjustment (BA) within a fixed window size to estimate the camera poses.

However, simply applying LEAP-VO in the camera estimation process is still unsatisfactory in most human-centric scenarios. The primary limitation stems from the dynamic movements of human subjects, which typically occupy a substantial portion of each image in human-centric videos. This dynamic presence introduces noise into the camera pose estimation in world coordinates, as the estimation process relies heavily on the relationship between the camera and the static environment. To address this issue, we propose a Masked LEAP-VO algorithm. Our approach involves inputting the image I_t and the human bbox at frame t into SAM [90] to generate a human mask. We then assign a visibility value of zero to points within the human mask, effectively excluding these trajectories from the BA process. For clarity, we denote S_{BA} as the window size of BA, $\hat{n}$ denotes the number of filtered point trajectories, and $w_{ij,\hat{n}}$ as the normalized weight based on confidence score and visibility. For estimating the camera poses $\mathbf{G} = \{\mathbf{R}, \mathbf{T}\}$ of orientation and translation, the reprojection loss function for BA can then be formulated as follows,

$$\mathbf{G} = \underset{\mathbf{G}, d_{i,\hat{n}}}{\arg\min} \sum_{i} \sum_{j \in |i-j| \leq S_{BA}} \sum_{\hat{n}} w_{ij,\hat{n}} ||\mathcal{F}(\mathbf{G}_i, \mathbf{G}_j, d_{i,\hat{n}}) - \Pi_{ij}(\mathbf{p}_{i,\hat{n}})||,$$

where $\mathcal{F}(\mathbf{G}_i, \mathbf{G}_j, d_{i,\hat{n}})$ denotes the point positions calculated by camera pose $\mathbf{G}$ at frame i and j with depth $d_{i,\hat{n}}$. $\Pi_{ij}(\mathbf{p}_{i,\hat{n}})$ denotes the position for project position of $\mathbf{p}_{i,\hat{n}}$ from frame i to j. Consequently, we obtain the camera rotation $\mathbf{R}_t$ and translation $\mathbf{T}_t$ from camera pose $\mathbf{G}_t$ at t.

Recovering human motion in world coordinates with estimated camera parameters. Given an input video, we feed the estimated camera parameters ($\mathbf{R}_t$ and $\mathbf{T}_t$) into the state-of-the-art motion recovering model, GVHMR [31],

$$\theta_t^w, \beta_t^w, \Gamma_t^w, \tau_t^w = \mathtt{GVHMR}(I_t, \mathbf{R}_t, \mathbf{T}_t). \tag{1}$$

Initialized human parameters $\theta_t^w, \beta_t^w, \Gamma_t^w, \tau_t^w$ and camera parameters $\mathbf{R}_t, \mathbf{T}_t$ will input to human motion alignment.

3.4. Aligning Human Motion Between Shots

Based on initialized world motion for each individual shot, the subsequent question is *how to merge discontinuous motions from different shots into a continuous motion sequence as a whole in world coordinates*. A straightforward solution is to align all motion sequences to the world coordinate system of the first shot. However, finding the correspondence between different shots is still under-explored and challenging. To resolve this issue, we decompose the motion parameters into camera-dependent and camera-independent ones. The former (Sec. 3.4.1) achieves alignment between shots via human orientation alignment based on camera calibration, whereas the latter (Sec. 3.4.2) is a trainable module to enhance the continuity of human motion sequence. These two key designs ensure a consistent motion sequence between frames when encountering shot transitions.

3.4.1 Aligning Human Orientations Between Shots

After obtaining the initial SMPL and camera parameters $\{\theta_t^i, \beta_t^i, \Gamma_t^i, \tau_t^i, \mathbf{R}_t^i, \mathbf{T}_t^i\}$ for each shot, directly concatenating motions between shots result abrupt changes of human poses and orientations. To address this issue, we introduce the *Orientation Alignment Module* (OAM), as shown

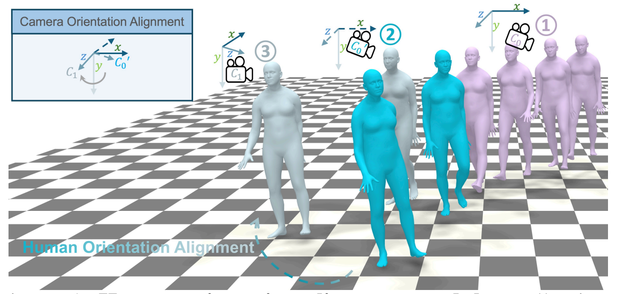

Figure 5. **Human orientation alignment module**. Following a shot transition after the foremost purple human mesh (shot ① captured by camera C_0), the unaligned (blue) and aligned (green) motions are captured as shot ② and shot "③" by camera C_0' and C_1, respectively. $C_0' = C_0$. To achieve human orientation alignment from shot "①" to "③", the camera rotation matrix from C_0' to C_1 is computed and applied as the offset of human orientation.

in Fig. 5, to align human orientations. As the whole motion sequence is continuous, we have the following assumption.

Assumption 1 *Human orientations and translations during the shot transition in world coordinates are continuous.*

To align the orientations between two frames with shot transition under Assumption 1, we decompose the human orientation with shot transitions in world coordinates as,

$$\mathtt{R}(\Gamma_{\text{world}}) = \mathbf{R}_{\delta_{\text{cam}}} \mathtt{R}(\Gamma_{\text{view}}), \tag{2}$$

where $\mathbf{R}_{\delta_{\text{cam}}}$ represents the camera rotation on the Y-axis between current t-th and previous $t-1$-th frame, Γ_{view} denotes the human orientation estimated by the current shot, and $\mathtt{R}(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{R}^9$ is the mapping from axis angle to rotation matrix. As Γ_{view} in current shot can be estimated independently, mentioned in Sec. 3.3, obtaining accurate Γ_{world} in Eq. (2) remains a key challenge to estimate the relative camera rotation $\mathbf{R}_{\delta_{\text{cam}}}$ between frames in shot transitions.

Estimating the relative camera pose $\mathbf{R}_{\delta_{\text{cam}}}$ between transition frames. Different from our approach of estimating camera pose in each shot (Sec. 3.3), we do not mask the human subject when estimating camera rotation $\mathbf{R}_{\delta_{\text{cam}}}$. Instead, we use human 2D KPTs as explicit feature matching. Specifically, we filter out unmatched keypoints based on their visibility and unaligned direction using RANSAC [91], effectively addressing camera pose estimation during shot transitions. This procedure is referred to as *Camera Calibration* (*a.k.a.* epipolar-geometry-based camera extrinsics estimation), and is detailed below.

In *Camera Calibration*, we assume that the human translations remain unchanged across the shot transition, implying that only the camera's orientation changes (*i.e.* Assumption 1). Consequently, we calculate the orientation offset by determining the change in camera orientation using camera calibration. We begin by extracting human 2D KPTs from two consecutive frames during the shot transition. Due to the shot transition, the visibility of 2D KPTs may vary,

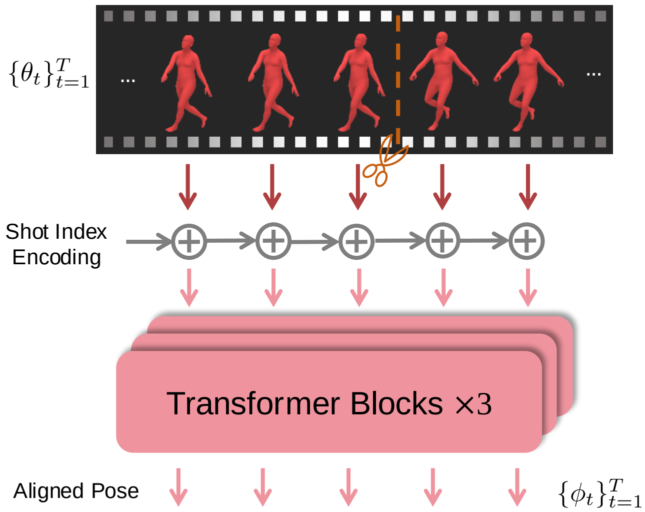

Figure 6. ***ms-HMR* Structure.** The initial human pose parameters θ across multiple video shots are input into a transformer with shot-index-based positional encoding. This enables *ms-HMR* to generate consistent human poses across all shots in the video.

e.g. occlusion in some shots. Therefore, we employ ED-Pose [92] to filter out invisible 2D KPTs between shot transition frames. Subsequently, RANSAC identifies matching 2D KPTs corresponding to the most possible camera rotation direction. These matched 2D KPTs facilitate the estimation of the aligned camera rotation $\mathbf{R}_{\delta_{\text{cam}}}$. The detailed estimation process is as follows.

We denote the detected 2D KPTs of two frames in the shot transition as $\mathbf{S}_1 = [(x_1^{(1)}, y_1^{(1)}), (x_1^{(2)}, y_1^{(2)}), \cdots, (x_1^{(N)}, y_1^{(N)})]^\top \in \mathtt{R}^{2\times N}$ and $\mathbf{S}_2 = [(x_2^{(1)}, y_2^{(1)}), (x_2^{(2)}, y_2^{(2)}), \cdots, (x_2^{(N)}, y_2^{(N)})]^\top \in \mathtt{R}^{2\times N}$. The essential matrix $\mathbf{E} = [\mathbf{T}]_\times \mathbf{R}$ should satisfy the following orthogonal property such that,

$$\mathbf{S}_1^\top \mathbf{E} \mathbf{S}_2 = \mathbf{0}. \tag{3}$$

Once $\mathbf{E}$ is obtained by solving Eq. (3), we enforce the rank-2 constraint on $\mathbf{E}$ through SVD decomposition and subsequently derive the aligned camera rotation $\mathbf{R}_{\delta_{\text{cam}}}$ between two frames (*cf*. Hartley *et al*. [93] for more details).

In summary, we reformulate the alignment problem of human orientation in shot transitions as estimating the relative camera rotation $\mathbf{R}_{\delta_{\text{cam}}}$ between frames. Accordingly, we obtain the camera rotation $\mathbf{R}_{\delta_{\text{cam}}}$ via camera calibration.

3.4.2 Aligning Human Poses Between Shots

In shot transition, video sequences recorded by two shots are often with various occlusions. However, unoccluded body parts in two shots can be complementary to each other for motion alignment. Thus, we introduce the *multi-shot HMR* (*ms*-HMR, *i.e*. $\mathtt{E}_M(\cdot)$) module to refine the whole motion sequence. As shown in Fig. 6, the *ms*-HMR is a Transformer encoder-like architecture, whose input and output

Dataset	Duration(s)	Videos	FPS	Max Length	Min Length	Shots
ms-Motion	23.7	600	30	1478	314	2, 3, 4

Table 1. Statistics of the *ms*-Motion dataset. By shots, we mean the number of shot transitions in a single video.

are the estimated global motion and the refined global motion, respectively. The process can be formulated as,

$$\phi_1, \phi_2, \cdots, \phi_T = \mathtt{E}_M(\theta_1, \theta_2, \cdots, \theta_T), \tag{4}$$

where ϕ_* denotes the refined motion of each frame. With this design, our method can adapt to diverse occlusions of human body brought by shot transitions.

3.5. Post-processing Module for Motion Integration

Trajectory and Foot Sliding Refiner. Inspired by Shin *et al*. [30], we introduce a bi-directional LSTM to recover foot-ground contact probabilities p_t^c, and root velocity v_t as,

$$p_t^c, v_t = \mathtt{LSTM}(\phi_1^m, \Gamma_1, \mathtt{F}(I_1), \phi_2^m, \Gamma_2, \mathtt{F}(I_2), \cdots, \phi_T^m, \Gamma_T, \mathtt{F}(I_T)), \tag{5}$$

where $\mathtt{F}(\cdot)$ denotes the image feature of each frame extracted by ViT [85]. Accordingly, the contact probabilities p_t^c, and velocity v_t are supervised by the ground-truth labels with MSE loss. Besides, we extend the trajectory refiner in WHAM [30] to improve the human trajectory estimation.

4. Benchmarking Multi-shot Motion Recovery

Dataset Construction. To create a multi-shot 3D human motion dataset, we introduce *ms*-Motion by processing existing public 3D human datasets with multiple camera settings and ground truth human and camera parameters, specifically AIST [64] and Human3.6M (H3.6M) [65]. In our construction pipeline, we randomly separate each original one-shot video into several clips. Then, we choose each clip from different shots and concatenate them together as one video recorded by multiple shots. For example, AIST provides each video with eight cameras `C0`, `C1`, ..., `C7` from different view point and we choose a video and split it into 5 clips at `t0`, `t1`, ..., `t4`. For frames in these separated clips, we choose frames shot by a random camera for each clip and combine five clips as one multi-shot video. Therefore, we construct a multi-shot version of AIST and H3.6M, which are named *ms*-AIST and *ms*-H3.6M subsets. Then we combine them and name this new dataset *ms*-Motion. The detailed statistics of *ms*-Motion are shown in Tab. 1. We do not compare with other existing 3D human datasets as they contain limited number of multi-shot videos.

Benchmark Evaluation Protocol. To evaluate the performance of our proposed methods on multi-shot videos, our target is to evaluate metrics for accurately reflecting the performance on videos with shot transitions. To this end, we use Root Orientation Error (*a.k.a*. ROE in *deg*) to measure the performance of the proposed method on human orientation alignment across different shots. Besides, we use Root

Dataset	Models	2-Shot				3-Shot				4-Shot			
		RTE↓	ROE↓	Jitter↓	Foot-Sliding↓	RTE↓	ROE↓	Jitter↓	Foot-Sliding↓	RTE↓	ROE↓	Jitter↓	Foot-Sliding↓
ms-AIST	SLAHMR [2023]	9.62	96.26	62.59	3.26	10.33	101.36	72.39	4.43	12.11	104.07	80.37	16.52
	WHAM [2024]	4.39	84.48	**25.24**	2.75	5.14	89.84	**24.06**	**2.99**	5.57	90.07	**26.29**	**3.62**
	GVHMR [2024]	6.20	96.58	34.87	7.65	7.55	99.69	34.46	9.42	8.96	104.53	35.67	9.78
	Ours	**2.56**	**69.23**	33.27	**2.66**	**3.64**	**67.71**	35.07	3.55	**4.55**	**70.31**	39.49	4.09
ms-H3.6M	SLAHMR [2023]	16.67	111.97	37.80	7.93	16.91	118.46	52.23	9.96	17.85	116.72	65.15	11.58
	WHAM [2024]	11.41	82.42	**18.40**	5.09	12.36	84.85	18.87	5.03	12.91	90.34	**18.40**	5.69
	GVHMR [2024]	6.94	81.93	18.45	8.80	85.25	58.26	18.36	10.62	9.12	91.63	19.47	10.65
	Ours	**3.65**	**53.39**	19.05	**4.17**	**5.33**	**58.26**	**17.35**	**4.62**	**6.20**	**61.22**	19.77	**5.12**

Table 2. **Quantitative comparison of different HMR methods on *ms*-Motion dataset.** We record the results for *ms*-AIST and *ms*-H3.6M separately. Our proposed method has achieved the best performance in RTE and ROE across *ms*-Motion among these methods.

Translation Error (*a.k.a.* RTE in m) to assess the performance of the proposed method on global trajectory recovery. Jitter ($\frac{10m}{fps^3}$) is also used to evaluate the stability of recovered human pose from multi-shot videos. We also include foot sliding (cm), the averaged displacement of foot vertices during contact with the ground, to assess the precision of recovered motion in the world coordinates [30].

5. Experiment

5.1. Datasets and Metrics

Evaluation Datasets. To evaluate the performance of our proposed pipeline for multi-shot videos, we use *ms*-Motion dataset and EMDB-1 dataset [77] with self-added noise for the evaluation of ablation study. For camera trajectory estimation, we use EMDB-1 and EMDB-2 split [77] as they contain the GT moving camera trajectory. Our self-created dataset contains 600 multi-shot videos, 42.7K frames, totaling 237 minutes. EMDB-1 split contains 17 video sequences totaling 13.5 minutes and EMDB-2 split contains 25 sequences totaling 24.0 minutes.

Evaluation Metrics. For shot detection we use *Recall*, *Precision* and *F1 Score* as evaluation metrics. For 3D human pose estimation-related tasks, we use ROE, RTE, jitter, and foot-sliding for evaluating the human motion recovery results on multi-shot videos. For the ablation study of our proposed pipeline, we evaluate the Procrustes-aligned Mean Per Joint Position Error (*a.k.a.* PA-MPJPE) and Per Vertex Error (*a.k.a.* PVE) as additional metrics besides previous mentioned ones. For camera pose estimation, we use absolute trajectory error (*a.k.a.* ATE) (m), Relative Pose Error (*a.k.a.* RPE) rotation (deg), and RPE translation (m).

5.2. Implementation Details

The *ms*-HMR, the trajectory, and foot sliding refiner are trained on the AMASS [82], 3DPW [83], Human3.6M [65], and BEDLAM [94] datasets, evaluate on EMDB and our *ms*-Motion. During training, we introduce random rotational noise (ranging from 0 to 1 radian) along the y-axis to the root pose Γ and random noise to the body pose θ at random positions to simulate the inaccuracies of pre-estimated human motions caused by shot transitions in multi-shot videos. This strategy enables the network to robustly recover smooth and consistent human motion from noisy initial parameters. The benchmark test results were obtained after training for 80 epochs on one NVIDIA-A100 GPU.

Methods	*ms*-Motion		
	Recall↑	Precision↑	F1 Score↑
Scenes Detect (SD) [87]	0.74	0.72	0.70
SD+Bbox Tracking (Bbox)	0.88	0.85	0.86
SD+Bbox+Pose Tracking	**0.96**	**0.88**	**0.92**

Table 3. **Comparison between difference shot detection algorithms.** We evaluate our shot transition detector on our proposed multi-shot video human motion dataset *ms*-Motion.

5.3. Main Results: Comparison of Global Human Motion Recovery Results on the Benchmark

We compare our proposed method *HumanMM* with several state-of-the-art HMR methods (SLAHMR [28], WHAM [30] and GVHMR [31]) on our proposed benchmark *ms*-Motion. As illustrated in Tab. 2, our proposed method has achieved the best performance for RTE and ROE through videos with all numbers of shots across *ms*-AIST and *ms*-H3.6M, indicating that our method reconstructs both the global human motion and orientations in the world coordinates more accurately and robustly. For the foot sliding metric, our method also performs as the best on *ms*-H3.6M across all numbers of shots.

5.4. Ablation Studies

Human-centric Scene Shot Boundary Detection Evaluation. To evaluate the performance of our proposed *Shot Transition Detector*, we test the algorithm on our proposed multi-shot human motion recovery benchmark and compare the output frame list of shot transitions with the ground truth (GT) of our dataset. As shown in Tab. 3, by applying the proposed finer granularity shot detection methods, the number of recall, precision, and F1 score all increases consistently. The combination of three steps (ScenesDetect, bbox tracking, and pose tracking) has achieved 0.96, 0.88,

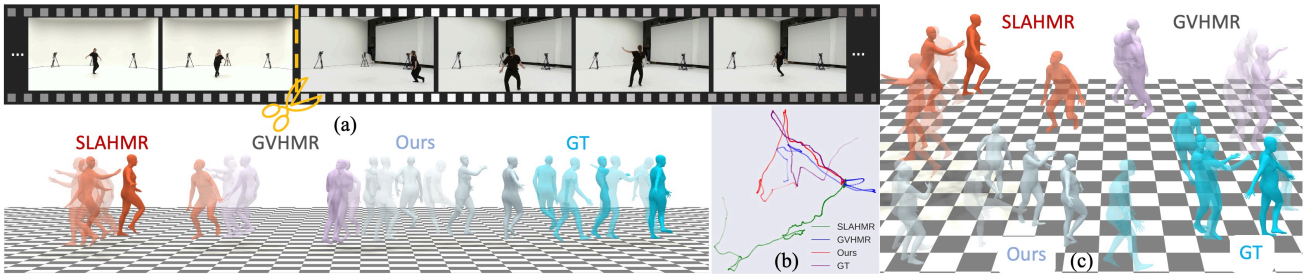

Figure 7. **Qualitative comparison of different HMR methods on *ms*-Motion dataset.** The side view of the rendered mesh for input mutli-shot video is shown in (a), while the top view is shown in (c). We also draw the comparison of the human trajectory as shown in (b). Our method is the most similar as GT in both rendered motion and trajectories among these methods.

Methods	PA-MPJPE↓	PVE↓	RTE↓	ROE↓	FS(foot sliding)↓
Baseline (Concat)	106.48	122.15	10.86	91.55	14.91
w/o *HumanMM*	78.24	85.77	3.89	50.63	3.54
w/o OAM	73.56	79.64	6.61	76.74	4.45
w/o traj. ref.	50.49	75.77	4.06	47.68	7.84
***HumanMM* (Ours)**	50.49	75.77	3.54	47.68	3.28

Table 4. **Ablation studies on different combinations of our modules.** We evaluate *HumanMM* on EMDB-1.

Methods	ATE↓	RPE trans↓	RPE rot↓
DPVO (w/o mask)	0.48	1.85	1.06
Masked DPVO	**0.48**	1.57	0.97
LEAP-VO (w/o mask)	0.50	0.93	0.97
Ours	0.51	**0.92**	**0.95**

Table 5. **Camera tracking results on EMDB 1 [77].** Our method has achieved $\sim 50\% \downarrow$ on RPE trans. than that of the original DPVO and perform the best in RPE rot.

Methods	ATE↓	RPE Trans.↓	RPE Rot.↓
DPVO (w/o mask)	**0.48**	1.07	1.26
Masked DPVO	0.50	0.86	1.21
LEAP-VO (w/o mask)	0.50	0.83	1.21
Ours	0.49	**0.83**	**1.19**

Table 6. **Camera tracking results on EMDB 2 [77].** Our method performs best. Besides, the masking operation is generally effective.

and 0.92 on the recall, precision, and F1 score, respectively, which indicates a comparable performance in shot boundary detection. Besides, as can be seen in the results, the latter two steps of shot detection contribute to the fine-grained final results significantly and jointly.

Key modules in the Proposed Method. We compare our methods with four variants on EMDB with noise dataset, as shown in Tab. 4, *ms*-HMR is the key component for the improvement in PA-MPJPE and PVE, which indicates a more accurate modeling of the whole motion sequence. This design serves as a recovery module to estimate some invisible body parts in some shots. Additionally, the orientation alignment module (*OAM*, in Sec. 3.4) is also a critical block for accurate human orientation estimation, indicated by the metric ROE. This module helps to model the global human motion between shots. For foot sliding, the results in Tab. 4 also show that the trajectory refiner (Sec. 3.5) in our method helps mitigate the foot sliding issue.

Comparison on Camera Trajectory Estimation. To evaluate the performance of our proposed camera trajectory estimation method **Masked LEAP-VO**, we evaluate the camera trajectory accuracy on EMDB 1 and EMDB 2. For more convenient comparison, we introduce two baselines, DPVO [80], which has been widely used in HMR methods such as WHAM [30] and GVHMR [31], and LEAP-VO [81]. To provide more intuition about the insights of masking dynamic humans in the video, we also implement a variant, Masked DPVO, by applying SAM at the patchify stage of DPVO to exclude patches containing human pixels. As shown in Tab. 5 and Tab. 6, compared with baseline methods, our key design of masking dynamic human subjects improves the result in both RPE Translation and RPE Rotation while maintaining competitive ATE. This result indicates the effectiveness of the design of masking dynamic human subjects in the process of camera trajectory estimation. Compared with the DPVO baseline, our method achieves $\sim 50\% \downarrow$ RPE translation on EMDB 1.

6. Conclusion and Discussion

Conclusion. In this paper, we introduce *HumanMM*, the first framework designed for human motion recovery from multi-shot videos in world coordinates. *HumanMM* addresses the challenges inherent in multi-shot videos by incorporating three key components: an enhanced camera trajectory estimation method called masked LEAP-VO, a human motion alignment module that ensures consistency across different shots, and a post-processing module for seamless motion integration. Extensive experiments demonstrate that *HumanMM* outperforms existing human motion recovery methods across various benchmarks, achieving state-of-the-art accuracy on our newly created multi-shot human motion dataset, *ms*-Motion.

Limitations and Future Work. While *HumanMM* represents an dvancement in human motion recovery from multi-shot videos in world coordinates, its performance may decline when faced with an excessive number of shot transitions. Despite these challenges, *HumanMM* provides a solid baseline for human motion recovery from multi-shot videos and can be employed in annotating *markerless* human motion datasets. Our newly introduced dataset, *ms*-Motion, offers a valuable benchmark for evaluating general human motion recovery methods in world coordinates, especially regarding their performance on multi-shot videos. Based on the proposed method, our future work aims to enlarge the related datasets for larger-scale motion databases.

References

[1] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, pages 20428–20437, 2022. 2

[2] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024. 2

[3] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, pages 580–597, 2022. 2

[4] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2024. 2

[5] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *3DV*, 2024.

[6] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *ICCV*, pages 20739–20749, 2023. 2

[7] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, pages 358–374, 2022. 2

[8] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022.

[9] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM SIGGRAPH*, 2022.

[10] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024.

[11] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, pages 414–423, 2022.

[12] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2022.

[13] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *NeurIPS*, pages 14959–14971, 2022.

[14] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2023.

[15] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, pages 9760–9770, 2023.

[16] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, pages 16010–16021, 2023.

[17] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pages 14730–14740, 2023.

[18] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024.

[19] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023.

[20] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *CVPR*, pages 2151–2162, 2023.

[21] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms, 2024.

[22] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *AAAI*, pages 7368–7376, 2024.

[23] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024.

[24] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*, 2024.

[25] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *ICML*, 2024. 2

[26] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 3

[27] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023. 2

[28] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 2, 3, 7

[29] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. *ECCV*, 2024. 2, 3

[30] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 6, 7, 8

[31] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *ACM SIGGRAPH Asia*, 2024. 2, 3, 4, 5, 7, 8

[32] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 3

[33] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and motion estimation from in-the-wild videos. In *3DV*, 2024. 2

[34] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 2

[35] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *RAS*, 109:13–26, 2018. 2

[36] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, pages 5915–5920, 2018.

[37] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018.

[38] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728, 2019.

[39] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *VR*, pages 1–10, 2021.

[40] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *ECCV*, 2024.

[41] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024.

[42] Jinpeng Liu, Wenxun Dai, Chunyu Wang, Yiji Cheng, Yansong Tang, and Xin Tong. Plan, posture and go: Towards open-world text-to-motion generation. *ECCV*, 2024.

[43] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *AAAI*, pages 2022–2030, 2024.

[44] Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *AAAI*, pages 6252–6260, 2024.

[45] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *ECCV*, 2024.

[46] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPRW*, pages 1911–1921, 2024.

[47] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, pages 457–469, 2024.

[48] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *CVPR*, pages 433–444, 2024.

[49] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. *ACM MM*, 2024.

[50] Jiaxu Zhang, Xin Chen, Gang Yu, and Zhigang Tu. Generative motion stylization of cross-structure characters within canonical motion space. In *ACM MM*, 2024.

[51] Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. Story-to-motion: Synthesizing infinite and controllable character animation from long text, 2023. 2

[52] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *NeurIPS*, 2022. 2

[53] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *CVPRW*, 2022.

[54] Gyeongsik Moon, Hongsuk Choi, Sanghyuk Chun, Jiyoung Lee, and Sangdoo Yun. Three recipes for better 3d pseudo-gts of 3d human mesh estimation in the wild. In *CVPRW*, 2023.

[55] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, pages 469–480, 2023. 2

[56] C.J. Bowen and R. Thompson. *Grammar of the Edit*. Focal Press, 2013. 2, 3

[57] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPRW*, 2024. 2

[58] 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. *NeurIPS*, 2023. 2

[59] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 2

[60] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 2, 3

[61] Peng Wu, Xiankai Lu, Jianbing Shen, and Yilong Yin. Clip fusion with bi-level optimization for human mesh reconstruction from monocular videos. In *ACM MM*, page 105–115, New York, NY, USA, 2023. Association for Computing Machinery.

[62] Kuan-Chieh Wang, Zhenzhen Weng, Maria Xenochristou, Joao Pedro Araujo, Jeffrey Gu, C Karen Liu, and Serena Yeung. Nemo: 3d neural motion fields from multiple video instances of the same action. In *CVPR*, 2023.

[63] Fabien Baradel, Thibault Groueix, Philippe Weinzaepfel, Romain Brégier, Yannis Kalantidis, and Grégory Rogez. Leveraging mocap data for human mesh recovery. In *3DV*, pages 586–595, 2021. 2

[64] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 3, 6

[65] 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 TPAMI*, 36(7):1325–1339, 2014. 3, 6, 7

[66] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6), 2017. 3

[67] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016.

[68] Anurag* Arnab, Carl* Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019.

[69] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. *STAR: Sparse Trained Articulated Human Body Regressor*, page 598–613. Springer International Publishing, 2020.

[70] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 3

[71] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[72] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[73] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[74] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 3

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*. Curran Associates, Inc., 2017. 3

[76] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, pages 13274–13285, 2022. 3

[77] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *ICCV*, 2023. 3, 7, 8

[78] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM TOG*, 40(4), 2021. 3

[79] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, pages 16558–16569. Curran Associates, Inc., 2021. 3

[80] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *NeurIPS*, 2023. 3, 8

[81] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *CVPR*, 2024. 3, 4, 8

[82] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 3, 7

[83] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3, 7

[84] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 4

[85] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4, 6

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

[87] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020. 4, 7

[88] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. `https://github.com/open-mmlab/mmtracking`, 2020. 4

[89] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *ECCV*, 2024. 4

[90] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 5

[91] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 5

[92] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *ICLR*, 2023. 6

[93] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 6

[94] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 7