COIN: Control-Inpainting Diffusion Prior for Human and Camera Motion Estimation

Jiefeng Li[1,2], Ye Yuan[1], Davis Rempe[1], Haotian Zhang[1], Pavlo Molchanov[1], Cewu Lu[2], Jan Kautz[1], and Umar Iqbal[1]

[1]NVIDIA [2]Shanghai Jiao Tong University
https://nvlabs.github.io/COIN/

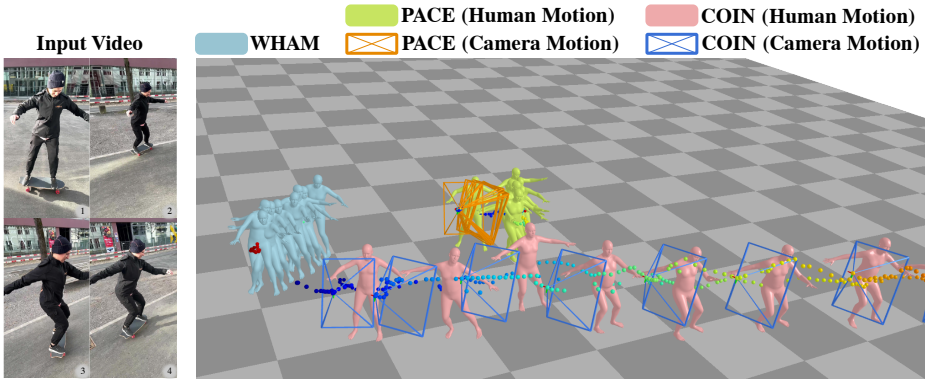

Fig. 1: Capturing global human and camera motion from a dynamic camera presents unique challenges. In the input video, a person is riding a skateboard – while the local body motion may remain relatively constant, the global position of the individual changes significantly. Current state-of-the-art methods such as PACE [41] and WHAM [81] fail catastrophically on such out-of-distribution motions. Our approach, COIN, gracefully handles such challenging cases, owing to our control-inpainting motion diffusion prior and novel human-scene relation loss.

Abstract. Estimating global human motion from moving cameras is challenging due to the entanglement of human and camera motions. To mitigate the ambiguity, existing methods leverage learned human motion priors, which however often result in oversmoothed motions with misaligned 2D projections. To tackle this problem, we propose COIN, a control-inpainting motion diffusion prior that enables fine-grained control to disentangle human and camera motions. Although pre-trained motion diffusion models encode rich motion priors, we find it non-trivial to leverage such knowledge to guide global motion estimation from RGB videos. COIN introduces a novel control-inpainting score distillation sampling method to ensure *well-aligned*, *consistent*, and *high-quality* motion from the diffusion prior within a joint optimization framework. Furthermore, we introduce a new human-scene relation loss to alleviate the scale ambiguity by enforcing consistency among the humans, camera, and scene. Experiments on three challenging benchmarks demonstrate the effectiveness of COIN, which outperforms the state-of-the-art methods in terms of global human motion estimation and camera motion estimation. As an illustrative example, COIN outperforms the state-of-the-art method by 33% in world joint position error (W-MPJPE) on the RICH dataset.

Keywords: Global Human Motion Estimation · Human Motion Prior · Score Distillation Sampling

1 Introduction

Recovering *global* human and camera motion from dynamic RGB videos is an important problem with many applications, such as animation, human-computer interaction, mixed reality, and robotics. However, it is a very challenging problem due to the entanglement of human and camera motion.

There are only a few works [41, 81, 99, 100] that try to address this problem. Earlier methods [47, 100] only focus on human motion and ignore the camera motion. Their core insight is that the global body motion is highly correlated with the local motion. Thus, they can use the local body movements to estimate the global orientation and trajectory with a regression model [100] or by combining them with physics constraints [47]. However, these regression models ignore the camera movements, so they fail to maintain consistency with the input video, whereas physics-based methods fail to model complex in-the-wild environments so are limited to controlled scenarios. Recent works [41, 99] try to jointly estimate the human and camera motion by exploiting learned motion priors [21, 76] and SLAM [66, 89, 90]. They try to constrain the human body motion in a low-dimensional latent space of a motion prior model, which results in reconstructed motions that are overly smooth and do not align well with video observations. Moreover, the optimization of the camera motion is only based on the global human motion from the motion prior. Hence, they fail catastrophically if the initial human motion predictions are significantly incorrect (as shown in Fig. 1).

More recently, Denoising Diffusion Models [24, 91] have emerged as a powerful family of generative models that can model high-quality data priors. Nonetheless, effectively leveraging the learned priors remains an ongoing challenge. Score Distillation Sampling (SDS) is commonly employed for this purpose [74], but we find that naive application of SDS also results in inconsistencies with the available observations (see Sec. 4.2). The root cause of this problem lies in the inconsistency of randomly sampled motions during SDS optimization. Without constraints, these motions may not align with observed evidence, leading to overly smoothed results that lack detail due to the mode-averaging effect.

In this work, we propose COIN, a hybrid **CO**ntrol-**IN**painting score distillation sampling method to address the aforementioned limitations of vanilla SDS. First, we use the partially observed human motion from the video as *control* signals to guide motion sampling. To tackle noisy observations which may be out of distribution for the prior, we propose a dynamic controlled sampling technique that iteratively refines the observed motions and updates the control signals to ensure effective distillation from the motion prior. Second, to further improve the consistency of the sampled motions, we also develop a novel *soft inpainting* strategy. We automatically identify the high-confidence regions of the initial predicted global motion from the video and use them as soft constraints during optimization. Concretely, we sample less confident regions from scratch using the

motion model, while the confident regions are only slightly refined. This ensures that the reconstructed motions do not deviate from the available observations. Our new SDS formulation is used to jointly optimize the human and camera motion by finding the most plausible solution that explains the observed evidence. Finally, to prevent catastrophic failure in cases where the initial body motion fails significantly, we propose a human-scene relation loss to consider the human-scene depth relations. This novel loss provides complementary information to the human motion prior by using local motion and scene features. It regularizes the camera scale by enforcing consistency among the human motion, camera motion, and scene features.

We benchmark our approach on the synthetic HCM [41] dataset and the real-world RICH [25] and EMDB [37] datasets. We demonstrate that our approach significantly outperforms the state-of-the-art methods in terms of human motion estimation and camera motion estimation. Overall, the contributions of this paper can be summarized as follows:

- We propose a novel control-inpainting motion prior specifically designed for global human motion estimation, which enhances score distillation sampling with dynamic control and soft inpainting to reconstruct well-aligned, consistent, and high-quality motions from video observations.
- We propose a new human-scene relation loss to resolve the scale ambiguity of the camera motion by enforcing consistency among the human motion, camera motion, and scene features.
- Our approach significantly outperforms the state-of-the-art methods in terms of human motion estimation and camera motion estimation on both synthetic and real-world datasets. In terms of global human motion estimation in world space, we outperform the state-of-the-art method PACE [41] by 44% and 33% on the HCM [41] and RICH [25] datasets, respectively. We also compare with the contemporary work WHAM [81] and outperform it by 49% and 7% on the RICH and EMDB datasets, respectively.

2 Related Work

Camera-Space Human Pose Estimation. Most existing works focus on root-relative local human pose estimation to bypass the difficulty in monocular depth estimation [1, 4, 6–8, 16, 32–34, 39, 42–46, 51, 54, 63, 64, 67, 68, 70, 76, 78, 84, 85, 88, 94, 97, 106, 110, 116, 119]. These methods ignore the position of the person in the camera coordinates. To overcome this limitation, recent methods estimate camera-space human poses by regression [28, 31, 48, 52, 62, 72, 75, 80, 93, 95, 107, 107, 109, 111] or optimization [59–61, 77, 108]. Physics-based constraints are widely used to ensure the plausibility of the estimated poses [10, 13, 27, 29, 80, 95, 105]. In addition to direct regression, heatmap-based representations have also been used to predict the absolute depths of multiple people [11, 87, 118]. A few methods improve absolution depth estimation by using predicted camera parameters [40, 49, 109] instead of the predefined focal length. Despite the promising results

for camera-space pose estimation, how to decouple the camera movement and estimate global human motions is still an open problem.

Monocular Global Human Pose Estimation. Recovering the global human motion from a monocular moving camera is challenging due to the entanglement of human and camera motions. To disentangle the camera movement, several methods use IMU sensors or pre-scanned environments to recover global human motions [17,20,57,69], which is impractical for large-scale adoption. Recent works use human motion priors [100] or physics-based constraints [47,55] to recover human motions from monocular videos, but do not consider background scene features, which limits performance on in-the-wild videos. Sun *et al*. [86] use optical flow as a motion cue to estimate the global motions. A contemporary work, WHAM [81], uses a lifting network to estimate global human motions from 2D keypoints and camera angular velocities. While these works can estimate accurate global human motions, they do not recover camera motions. To explicitly recover the camera motion, Liu *et al*. [52] use SLAM and convert the local pose from the camera to global coordinates. BodySLAM [22] jointly optimizes the human and camera motion using features of both humans and scenes. Along this line, SLAHMR [99] and PACE [41] use SLAM to initialize camera motions and optimize the camera using human motion priors [21,76]. However, these methods rely on the human motion priors to regularize the camera motion, which may lead to inaccurate camera motion when the human motion is not well initialized (as shown in Fig. 1). Such wrong camera trajectories will further affect the optimization of human motions. In contrast, our approach relies on the consistency among the local human motion, scene features, and the camera for optimization, which provides information that complements the human motion priors and enables accurate estimation of both human and camera motions.

Human Motion Priors. There are a significant amount of approaches proposed to study 3D human dynamics for motion prediction and synthesis [2,3,5,12,15,18,19,23,30,36,38,50,58,71,73,92,98,101–103]. These learned human motion priors are used to help resolve pose ambiguity [21,39,76,115] in human pose estimation. Recently, diffusion models [82] have also been used as priors for motion synthesis and infilling [26,35,91,104,113]. RoHM [114], adopts motion diffusion model to recover human motions from noisy and occluded input data. Xie *et al*. [96] use spatial control signals to guide motion generation. They focus on generating realistic human motion given clean spatial constraints. Müller *et al*. [65] build a diffusion model to learn the joint distribution over the poses of two people. They use the SDS loss to guide the generation of static poses. In contrast, we focus on dynamic human motions. We find adopting SDS directly for temporal human motions encounters the inconsistency issue. Therefore, to distill knowledge from the motion diffusion model, we propose a novel control-inpainting SDS to generate high-quality and consistent motion that aligns with observed evidence.

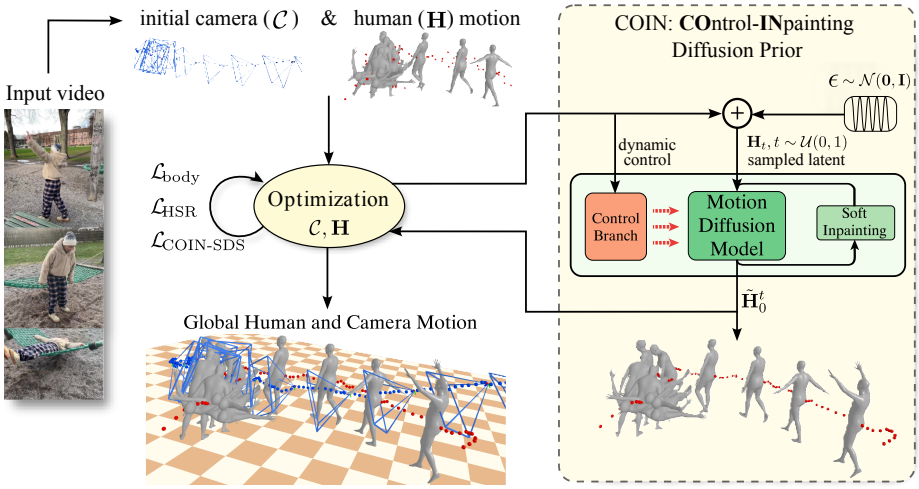

Fig. 2: Overview. Given a video with a moving camera, we recover the global human motion $\mathbf{H}$ and camera motion $\mathcal{C}$ using an iterative optimization framework. We propose a novel Control-Inpainting SDS loss ($\mathcal{L}_{\text{COIN-SDS}}$) to leverage motion diffusion models as a prior. COIN-SDS is designed such that the sampled motions from the motion prior are consistent with video observations. We achieve this by controlling and constraining the sampling process of the motion diffusion model through novel control and soft-inpainting branches. We also propose a novel human-scene relation loss ($\mathcal{L}_{\text{HSR}}$) to encourage consistency among the human motion, camera motion, and scene features.

3 Method

The overall framework of COIN is illustrated in Fig. 2. Given an in-the-wild RGB video with T frames captured by a dynamic camera, our goal is to estimate both the global human motion $\mathbf{H} = \{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \ldots, \mathbf{h}^{(T)}\}$ and the camera motion $\mathcal{C} = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(T)}\}$ in a global world coordinate system. We use off-the-shelf 3D human pose and shape estimation method HybrIK [48] to obtain per-frame initial SMPL parameters in the camera space and DROID-SLAM [89] to obtain the initial per-frame camera-to-world transforms. We convert the local human motion to the world coordinates with the estimated camera. However, because the camera trajectories from SLAM are up to an unknown scale, the initial global human motion will abnormally drift and float in the world space. To resolve the scale ambiguity and place the person in the correct global position, we jointly optimize the human and camera motion to minimize the discrepancy between the observed evidence and the estimated motion, while maintaining the plausibility of the human motion with a diffusion prior through the proposed control-inpainting SDS.

Motion Representation. The camera motion is represented by the trajectory $\mathcal{C} = \{(\mathbf{R}^{(i)}, \mathbf{t}^{(i)})\}_{i=1}^{T}$, where $[\mathbf{R}^{(i)}, \mathbf{t}^{(i)}]$ is the camera pose at the i-th frame, consisting of the rotation matrix $\mathbf{R}^{(i)} \in \mathbb{R}^{3\times3}$ and the translation vector $\mathbf{t}^{(i)} \in \mathbb{R}^3$. The human motion is represented by the human trajectory $\mathbf{H} = \{\mathbf{h}^{(i)}\}_{i=1}^{T}$,

where $\mathbf{h}^{(i)} = [\tau^{(i)}, \Phi^{(i)}, \theta^{(i)}, f^{(i)}, \beta]$ is the human pose at the i-th frame, consisting of the global translation $\tau^{(i)} \in \mathbb{R}^3$, global orientation $\Phi^{(i)} \in \mathbb{R}^3$, body pose parameters $\theta^{(i)} \in \mathbb{R}^{23\times3}$, foot contact labels $f \in \{0,1\}^4$, and the body shape parameters $\beta \in \mathbb{R}^{10}$. We use the SMPL model [53] to represent the human pose and shape. Before introducing our approach, we first revisit the formulation and drawbacks of SDS in Sec. 3.1. Then we introduce the proposed control-inpainting SDS in Sec. 3.2. Finally, we present the global optimization pipeline with the proposed human-scene interaction loss in Sec. 3.3.

3.1 Revisiting SDS

Score Distillation Sampling (SDS) was first introduced to distill 3D assets from pre-trained 2D text-to-image diffusion models [74]. It exploits the knowledge from the diffusion models by seeking modes for the conditional distribution in the DDPM latent space to optimize the 3D scene representation. Similarly, we can optimize global human motion by distilling knowledge from a pre-trained motion diffusion model.

Given an global human motion $\mathbf{H}$, the marginal distribution of noisy latent $\mathbf{H}_t$ at timestep $t \in \mathcal{U}(0,1)$ is defined as:

$$q(\mathbf{H}_t|\mathbf{H}) = \mathcal{N}(\mathbf{H}_t; \sqrt{\bar{\alpha}_t}\mathbf{H}, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t \in (0,1)$ is a hyperparameter controlled by the variance schedule of the diffusion model. SDS adopts the pre-trained diffusion model $\mathcal{D}_\phi(\mathbf{H}_t, t, y)$, which takes in $\mathbf{H}_t$ and is used to model the conditional density of the human motion, where ϕ are the parameters of the diffusion model and y is the condition. Then, SDS aims to distill global human motion $\mathbf{H}$ via seeking modes of the learned conditional density, which can be achieved by a weighted denoising score matching objective:

$$\min_{\mathbf{H}} \mathcal{L}_{\text{SDS}} := \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\omega(t) \|\boldsymbol{\epsilon}_\phi^t - \boldsymbol{\epsilon}\|_2^2 \right], \tag{2}$$

where $\boldsymbol{\epsilon}_\phi^t$ is the predicted denoising direction from the diffusion model, $\mathbf{H}_t \sim q(\mathbf{H}_t|\mathbf{H})$ is sampled using the reparameterization trick, $\boldsymbol{\epsilon}$ is the corresponding sampled noise, and $\omega(t)$ is a weighting function that depends on the timestep t.

To clearly review the effect of SDS, we can reparameterize Eq. 2 as:

$$\min_{\mathbf{H}} \mathcal{L}_{\text{SDS}} := \mathbb{E}_t \left[\frac{\omega(t)\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} \left\| \mathbf{H} - \hat{\mathbf{H}}_0^t \right\|_2^2 \right], \tag{3}$$

where

$$\hat{\mathbf{H}}_0^t = \frac{\mathbf{H}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\phi^t}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

Based on this reparameterization, we can see that the SDS objective is to minimize the discrepancy between global human motion $\mathbf{H}$ and the denoised global human motion $\hat{\mathbf{H}}_0^t$ from the motion diffusion model in a single step. The denoised

motion $\hat{\mathbf{H}}_0^t$ serves as the *pseudo ground truth*. However, at each optimization step, we randomly sample t and $\boldsymbol{\epsilon}$ to generate the noisy latent $\mathbf{H}_t$, and we found the pre-trained diffusion model is sensitive to the input. Minor fluctuations in the input latent would substantially change the denoised motion, which leads to inconsistency in $\hat{\mathbf{H}}_0^t$ across different time steps.

Although randomness can help generate diverse plausible motions to infer occluded regions and unknown information, we do not need it for well-observed regions, such as simple body poses in a clean background. Such randomness in the denoising steps makes the generated $\hat{\mathbf{H}}_0^t$ difficult to align with the local 2D observations and results in wrong global human motion. Moreover, this pseudo ground truth $\hat{\mathbf{H}}_0^t$ is generated from only a single denoising step, where the diffusion models may not produce high-quality motions, resulting in foot sliding and floating. Although sampling with a smaller timestep t can alleviate these issues, the initial motion is usually inaccurate and the denoiser is not able to remove artifacts with a small t. To exploit the knowledge of the motion diffusion model and denoise the initial motion, we must allow the SDS to sample with a larger timestep t while maintaining high quality, consistency, and alignment with the local 2D observations.

3.2 Control-Inpainting SDS

The limitations of SDS originate from the randomness and inconsistency of the denoised motion $\hat{\mathbf{H}}_0^t$, which serves as the pseudo ground truth in the objective function. To overcome this issue, we propose a novel **CO**ntrol-**IN**painting SDS (COIN-SDS) to generate *high-quality* and *consistent* pseudo-ground-truth motions. Our solution has three key ingredients (shown in Alg. 1). First, to achieve *high-quality* motions, we seek to produce the pseudo ground truth with multiple DDIM denoising steps. Second, to encourage *consistent* motions, we propose to use partially observed evidence from the video as a control signal to dynamically guide the diffusion model and align the generated motions with the observations. Third, to further align the motion with observed regions, we propose a soft inpainting strategy within the denoising process.

Multiple Denoising Steps. Intuitively, to obtain high-quality pseudo ground truth for SDS, we can replace the single-step denoised motion $\hat{\mathbf{H}}_0^t$ with a multi-step one $\tilde{\mathbf{H}}_0^t := \tilde{\mathbf{H}}_0$, following the multi-step DDIM denoising process [83]:

$$\tilde{\mathbf{H}}_{t-\Delta t} = \sqrt{\bar{\alpha}_{t-\Delta t}} \cdot \hat{\mathbf{H}}_0^t + \sqrt{1-\bar{\alpha}_{t-\Delta t}} \cdot \boldsymbol{\epsilon}_\phi^t, \tag{5}$$

until $\tilde{\mathbf{H}}_0 = \tilde{\mathbf{H}}_{t-\Delta t}$ is obtained. By replacing $\hat{\mathbf{H}}_0^t$ in Eq. 3 with $\tilde{\mathbf{H}}_0$, we can obtain a new objective for SDS:

$$\min_{\mathbf{H}} \mathcal{L}_{\text{SDS}} := \mathbb{E}_t \left[\frac{\omega(t)\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} \left\| \mathbf{H} - \tilde{\mathbf{H}}_0 \right\|_2^2 \right]. \tag{6}$$

Although the multi-step denoising process can produce high-quality pseudo ground truth, it is computationally expensive to perform multiple denoising steps during

Algorithm 1: COIN-SDS

Input: Latest human motion $\mathbf{H}$, confidence score $\mathbf{S}$, visible mask $\mathbf{M}$
Output: $\mathcal{L}_{\text{COIN-SDS}}$

Sample: $\tilde{\mathbf{H}}_t \sim \mathcal{N}(\mathbf{H}_t; \sqrt{\bar{\alpha}_t}\mathbf{H}, (1-\bar{\alpha}_t)\mathbf{I})$, $t \sim \mathcal{U}(0,1)$;
for $\bar{t} = [t, t-\Delta t, \ldots, \Delta t]$ **do** // multi-step DDIM denoising
  $\tilde{\mathbf{H}}_0^{\bar{t},\text{known}} \leftarrow \mathbf{H}$;
  $\tilde{\mathbf{H}}_0^{\bar{t},\text{unknown}} \leftarrow \mathcal{D}_{\phi,\phi_c}(\tilde{\mathbf{H}}_{\bar{t}}, \bar{t}, \mathbf{H} \odot \mathbf{M})$; // controlled denoising
  $\tilde{\mathbf{M}} \leftarrow w(\bar{t}) * \mathbf{S} \odot \mathbf{M}$;
  $\tilde{\mathbf{H}}_0^{\bar{t}} \leftarrow \tilde{\mathbf{M}} \odot \tilde{\mathbf{H}}_0^{\bar{t},\text{known}} + (1-\tilde{\mathbf{M}}) \odot \tilde{\mathbf{H}}_0^{\bar{t},\text{unknown}}$; // soft inpainting
  $\boldsymbol{\epsilon}_\phi^{\bar{t}} \leftarrow \frac{\tilde{\mathbf{H}}_{\bar{t}} - \sqrt{\bar{\alpha}_{\bar{t}}}\tilde{\mathbf{H}}_0^{\bar{t}}}{\sqrt{1-\bar{\alpha}_{\bar{t}}}}$;
  $\tilde{\mathbf{H}}_{\bar{t}-\Delta t} \leftarrow \sqrt{\bar{\alpha}_{\bar{t}-\Delta t}} \cdot \tilde{\mathbf{H}}_0^{\bar{t}} + \sqrt{1-\bar{\alpha}_{\bar{t}-\Delta t}} \cdot \boldsymbol{\epsilon}_\phi^{\bar{t}}$; // update latent motion
$\mathcal{L}_{\text{COIN-SDS}} = \frac{\omega(t)\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} \left\| \mathbf{H} - \tilde{\mathbf{H}}_0 \right\|_2^2$

optimization, which limits the practicality of increasing the number of denoising steps. In our experiments, we find that using 10 denoising steps is sufficient to produce high-quality pseudo ground truth.

Dynamic Controlled Sampling. To generate consistent motions that are aligned with the observed evidence, we propose to attach a control branch ϕ_c to the pre-trained diffusion model $\mathcal{D}_\phi$ to guide the motion generation. Given a latent motion $\tilde{\mathbf{H}}_t$, control signal $\mathbf{c}$, and the visible mask $\mathbf{M}$, we train a controlled denoiser $\mathcal{D}_{\phi,\phi_c}$ to generate intermediate denoised motion $\tilde{\mathbf{H}}_0^t$ for DDIM denoising:

$$\tilde{\mathbf{H}}_0^t = \mathcal{D}_{\phi,\phi_c}(\tilde{\mathbf{H}}_t, t, \mathbf{c} \odot \mathbf{M}), \tag{7}$$

where $\mathbf{c}$ and $\mathbf{M}$ are the same size as the motion, $\mathbf{M}$ is a binary mask with ones in observed pose dimensions, and $\odot$ denotes the element-wise multiplication. During training, we synthesize noise and occlusions by randomly adding Gaussian noise to the control signal $\mathbf{c}$ and randomly masking the pose and trajectory dimensions in $\mathbf{M}$. Details are provided in the appendix.

When using the denoiser in the optimization stage, instead of always using the initial noisy estimation from HybrIK [48] as the fixed control signal, we propose to use a *dynamic control strategy*. Specifically, we use the optimized human motion from the previous iteration as the control signal, *i.e.*, $\mathbf{c} = \mathbf{H}$. This strategy prevents performance degradation due to inaccurate initializations. A better control signal can guide the model to generate a more plausible pseudo motion, which in turn helps to optimize the global human motion and provides a control signal that better aligns with the input videos. Such self-evolving control signals can help to generate well-aligned global human motions.

The pre-trained motion diffusion model adopts a transformer encoder structure like [91]. We follow ControlNet [112] to encode the control signals and guide the denoiser output. We create a trainable copy of 4 encoding blocks of the pre-trained motion diffusion model followed with zero convolutions. The input to the control branch is the concatenation of the latent and the control signals. The

AMASS [56] dataset is used to train the motion diffusion model. The finetuning of the controlled denoiser is computationally efficient since the pre-trained branch is frozen and only the control branch is trained. See the appendix for more details on the architecture and training settings.

Soft Inpainting. While guiding motion generation encourages outputs to align with the conditions, it is often not strong enough. We further seek to improve the consistency by masking the known regions and inpainting the unknown regions. Given a binary mask $\mathbf{M}$ that indicates the observed and unobserved regions, we can use the diffusion model to generate the inpainted motion $\tilde{\mathbf{H}}_0^t$ following the DDIM denoising process:

$$\tilde{\mathbf{H}}_0^{t,\text{known}} = \mathbf{H}, \tag{8a}$$

$$\tilde{\mathbf{H}}_0^{t,\text{unknown}} = \mathcal{D}_{\phi,\phi_c}(\tilde{\mathbf{H}}_t, t, \mathbf{H} \odot \mathbf{M}), \tag{8b}$$

$$\tilde{\mathbf{H}}_0^t = \mathbf{M} \odot \tilde{\mathbf{H}}_0^{t,\text{known}} + (1 - \mathbf{M}) \odot \tilde{\mathbf{H}}_0^{t,\text{unknown}}. \tag{8c}$$

Thus, the known regions are overwritten with the observations, while the unknown regions are sampled from the diffusion model. However, the above formulation keeps the observed parts unchanged during the denoising process. In practice, the observed parts can be noisy and not perfect. We still want the diffusion model to refine the observed parts but not change them significantly.

Here, we present a soft inpainting strategy to infill the unobserved regions while refining the observed regions by dynamically reweighting the denoised direction from the diffusion model. Specifically, instead of using a binary mask, we adopt a continuous mask $\tilde{\mathbf{M}}$ depending on both the confidence score of the observations $\mathbf{S}$ and the denoising time step t:

$$\tilde{\mathbf{M}} = w(t) * \mathbf{S} \odot \mathbf{M}, \tag{9}$$

where we set $w(t) = \max(0, \frac{t-0.5}{0.5})$ to linearly decrease the weight of the observations as the denoising time step decreases. As the time step decreases, the denoising process will be more deterministic and the model will be more certain of the generated motions.

Combining the three components of our solution, the final objective for COIN-SDS is formulated as:

$$\min_{\mathbf{H}} \mathcal{L}_{\text{COIN-SDS}} := \mathbb{E}_t \left[\frac{\omega(t)\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} \left\| \mathbf{H} - \tilde{\mathbf{H}}_0(\mathbf{H}, \mathbf{M}, \mathbf{S}, t) \right\|_2^2 \right]. \tag{10}$$

We summarize COIN in Alg. 1.

3.3 Global Optimization

Here we present the overall optimization pipeline for the joint estimation of global human and camera motion with the proposed COIN-SDS loss. Note that we use SLAM to initialize the camera poses, which is scale-ambiguous. Therefore,

we need to jointly optimize the camera scale s with the human and camera motions. Furthermore, the SLAM method assumes the camera in the first frame to be at the origin. To put the human motion in the correct positions, we follow PACE [41] and also optimize the camera height h_0 and the orientation R_0 for the first frame. The global human motions are initialized by the estimated local motions and the camera poses. The overall optimization objective is:

$$\min_{\mathbf{H},\mathcal{C},s,h_0,R_0,\beta} \mathcal{L}_{\text{body}} + \mathcal{L}_{\text{COIN-SDS}} + \mathcal{L}_{\text{HSR}}, \tag{11}$$

where

$$\mathcal{L}_{\text{body}} = \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{3D}} + \mathcal{L}_{\beta} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{contact}}. \tag{12}$$

$\mathcal{L}_{\text{2D}}$ measures the 2D reprojection error between the projected 2D body joints of the estimated human motion and the detected 2D keypoints from an off-the-shelf 2D joint detector [9]. $\mathcal{L}_{\text{3D}}$ measures the distance between the estimated local 3D joints and the detected 3D joints from an off-the-shelf 3D joint detector [79]. $\mathcal{L}_{\beta}$ is the shape regularization loss. $\mathcal{L}_{\text{smooth}}$ is the temporal smoothness loss. $\mathcal{L}_{\text{contact}}$ is the foot contact loss to encourage zero velocities for contact joints. The contact labels are obtained from the pseudo ground truth motion from COIN-SDS. Please refer to the appendix for more details.

Human-scene Relation Loss. The camera trajectories recovered from SLAM are scale-ambiguous. Previous works [41, 99] optimize the camera scale by projecting the global human motion to the camera space using the camera poses and minimizing the reprojection error. However, such a method entirely relies on the global human motions, which is in turn affected by the camera scale. If the human motion is not initialized well, the camera scale will also be inaccurate. To solve this problem, instead of the global human motions, we propose a new human-scene relation loss that uses the depth relation between the human and scene in the camera space, which disentangles the effect of the camera itself.

Specifically, we use the point cloud of the scene recovered by SLAM as a constraint. First of all, the scale of the scene point cloud is the same as the camera scale, so optimizing the scene scale is equivalent to optimizing the camera scale. Second, the scene points that are projected onto the visible vertices of the body mesh should be occluded by the person. Otherwise, the corresponding body parts are invisible. Therefore, we can constrain the depth of the occluded scene points to be larger than the depth of the human body vertices. While finding the corresponding body vertices for each scene point is time-consuming, we propose to use the depth of its nearest body joint as a proxy. Given the scene point cloud $\mathcal{P}$ and the camera scale s, the human-scene relation loss is formulated as:

$$\mathcal{L}_{\text{HSR}} = -\frac{1}{|\mathcal{P}|} \sum_{i=1}^{T} \sum_{p \in \mathcal{P}^*} \min(0, \mathcal{T}^{(i)}(p)_z - j^{(i)}(p)_z) \cdot \mathbb{1}(\mathcal{T}^{(i)}(p) \text{ is invisible}), \tag{13}$$

where $\mathcal{P}^* = \mathcal{P} * s$ is the scaled point cloud of the scene, $\mathcal{T}^{(i)}(p) = \mathbf{R}^{(i)}p + \mathbf{t}^{(i)}$ is the transformed point in the i-th frame, $j^{(i)}(p)$ is body joint that has the nearest

2D projection to the scene point p in the i-th frame, and z denotes the depth of a given point. If the depth order is correct, *i.e.*, $\mathcal{T}^{(i)}(p)_z - j^{(i)}(p)_z > 0$, the loss is zero. The proposed human-scene relation loss uses the relation between the local motions and the scene to alleviate the scale ambiguity. The depth relation regularizes consistency among humans, cameras, and scenes.

4 Experiments

Datasets. We perform experiments on three human motion datasets. First is the real-world dataset RICH [25]. We follow previous works [41,81,100] to assess the performance of global human motion estimation using this dataset. The second one is EMDB [37]. We follow previous works [81] to evaluate on a subset of EMDB for which they provide ground truth global motion with dynamic cameras. The third dataset is HCM [41], a synthetic dataset. Compared to real-world datasets, HCM contains more challenging camera motions. We follow previous work [41] to evaluate the global human motion and the camera motion using this dataset.

Metrics. We report various metrics for both human and camera motion. For human motion, standard metrics W-MPJPE and WA-MPJPE are used to evaluate global motion, while PA-MPJPE evaluates local motion. We also include an ACCEL metric to measure the joint acceleration error. For evaluation on EMDB, we follow previous work [81] to split sequences into smaller chunks of 100 frames and align each output segment with the ground-truth data using the first two frames W-MPJPE$_{100}$ or the entire segment WA-MPJPE$_{100}$. Root Orientation Error (ROE in `deg`) and Root Translation Error (RTE in m) evaluate the error over the entire trajectory after aligning with the initial camera pose.

For camera motion, we report the average translation error after scale alignment (ATE), without scale alignment (ATE-S), and the camera acceleration error (CAM ACCEL). ATE-S more accurately reflects inaccuracies in the captured scale of the camera.

Baselines. As discussed in Sec. 2, there are different ways to use the pre-trained motion diffusion model as a motion prior. Here, we summarize the three main solutions and compare them with COIN in Sec. 4.2. (**1**) Guided Sampling, which embeds analytical guidance within the denoising procedure using objective functions, such as 2D projection and foot contact consistency. This does not suit our task because the camera trajectories are also unknown and guided gradients from the wrong camera will lead to unrealistic human motions. We need to optimize the human motion and camera motion simultaneously. (**2**) Noise Optimization, which represents the motion as latent noise and directly optimizes it. This is similar to other motion priors such as VAE [21]. At each optimization step, we need to calculate the gradients of the latent w.r.t. the generated motion, which is computationally expensive and we find the performance is not good enough. (**3**) Vanilla SDS, which removes our design and directly uses SDS for optimization.

Table 1: Global human motion estimation on the RICH dataset.

Method	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	W-RJE ↓	ACCEL ↓
HybrIK [48] + SLAM [89]	46.7	1073.1	404.4	1166.2	20.2
GLAMR [100]	79.9	653.7	365.1	646.6	107.7
SLAHMR [99]	52.5	571.6	323.7	400.5	9.4
WHAM [81]	46.2	497.6	272.7	478.2	**6.7**
PACE [41]	49.3	380.0	197.2	370.8	8.8
Guided Sampling	132.8	1384.6	502.2	1440.9	24.2
Noise Optimization	66.7	414.8	195.3	429.2	8.4
Vanilla SDS	78.8	1453.5	497.2	1458.0	12.7
COIN w/o Controlled Sampling	44.0	825.0	291.8	848.5	10.8
COIN w/o Dynamic Control	49.5	293.8	180.6	299.1	8.6
COIN w/o Soft Inpainting	47.6	325.8	196.0	324.9	9.6
COIN w/o $\mathcal{L}_{\text{HSR}}$	43.6	273.0	176.3	281.1	8.2
COIN	**42.9**	**254.5**	**169.5**	**249.9**	7.5

Table 2: Global human motion estimation on the EMDB dataset.

Method	PA-MPJPE ↓	W-MPJPE$_{100}$ ↓	WA-MPJPE$_{100}$ ↓	RTE ↓	ROE ↓
HMR2.0 [14] + DPVO [90]	49.6	2320.9	662.9	17.5	44.4
GLAMR [100]	56.0	756.1	286.2	16.7	74.9
TRACE [86]	58.0	2244.9	544.1	18.9	72.7
SLAHMR [99]	61.5	807.4	336.9	13.8	67.9
WHAM [81]	41.9	439.2	166.1	8.4	36.3
COIN	**32.7**	**407.3**	**152.8**	**3.5**	**34.1**

4.1 Comparison with State-of-the-Art Methods

Human Motion Estimation. We compare COIN against state-of-the-art methods on the RICH, EMDB, and HCM datasets. Quantitative results are shown in Tabs. 1, 2, and 3. We observe that COIN significantly outperforms the state-of-the-art methods on all datasets. On the RICH dataset, COIN outperforms the state-of-the-art method, PACE [41], by **125.5** mm in terms of W-MPJPE, showing **33.0**% relative improvement. On the EMDB and HCM datasets, COIN shows **29.1** mm and **109.0** mm improvement in terms of W-MPJPE, respectively. Qualitative comparisons with state-of-the-art methods, PACE [41] and WHAM [81], are shown in Figs. 1 and 3. More qualitative results are shown in the supplementary video.

COIN not only improves global motion, but also improves local body motion. In terms of PA-MPJPE, COIN shows **3.3** mm, **9.2** mm, and **19.8** mm improvement on the RICH, EMDB, and HCM datasets, respectively. This demonstrates that COIN is able to distill high-quality motion priors from the diffusion model and help both local and global motion estimation. Regarding the joint acceleration error, COIN is 0.8 mm/s^2 higher than WHAM [81]. Note that the joint acceleration error reflects the smoothness. Human motion can be over-smoothed but not accurate, hence it is important to look at acceleration error in conjunction with other metrics.

Table 3: Global human motion estimation on the HCM dataset.

Method	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	W-RJE ↓	ACCEL ↓
HybrIK [48] + SLAM [89]	67.6	1137.3	780.3	1100.9	51.3
GLAMR [100]	86.0	1977.6	653.8	1958.0	33.4
SLAHMR [99]	69.9	888.9	483.5	862.2	14.9
PACE [41]	65.3	861.2	478.3	839.5	16.7
WHAM [81]	47.9	588.9	279.3	579.2	13.1
COIN	**45.5**	**479.9**	**212.1**	**470.7**	**10.1**

Table 4: Camera motion estimation on the HCM dataset.

Method	ATE ↓	ATE-S ↓	CAM ACCEL ↓
HybrIK [48] + SLAM [89]	155.8	1670.7	17.1
GLAMR [100]	1295.2	1714.6	282.9
SLAHMR [99]	155.8	506.5	17.6
PACE [41]	137.5	459.7	16.2
Guided Sampling	335.6	992.4	15.6
Noise Optimization	206.0	500.9	12.3
Vanilla SDS	306.4	656.4	13.7
COIN w/o Controlled Sampling	299.6	553.8	14.0
COIN w/o Dynamic Control	149.8	397.7	**11.3**
COIN w/o Soft Inpainting	167.7	423.8	11.4
COIN w/o $\mathcal{L}_{\text{HSR}}$	147.8	402.1	12.0
COIN	**135.3**	**385.9**	**11.3**

Camera Motion Estimation. We further evaluate the performance of COIN on camera motion estimation on the HCM dataset. Quantitative results are shown in Tab. 4. COIN substantially surpasses the state-of-the-art camera motion estimation methods. Specifically, COIN reduces the absolute camera translation error, ATE-S by **73.8** mm. This demonstrates that COIN is able to disentangle human and camera motions and accurately estimate the camera motion.

4.2 Ablation Study

In this section, we conduct ablation studies on the RICH and HCM datasets to evaluate the impact of each component on human and camera motions, respectively. More comparisons on other datasets are provided in the appendix.

Baselines with Motion Diffusion Model. We first compare the aforementioned baselines with COIN. To use guided sampling in our tasks, we jointly update the camera poses using the gradients from the objective function during denoising. Quantitative results of human and camera motion estimation are shown in Tabs. 1 and 4, respectively. We observe that COIN outperforms all the baselines in terms of all metrics. As expected, guided sampling shows a terrible

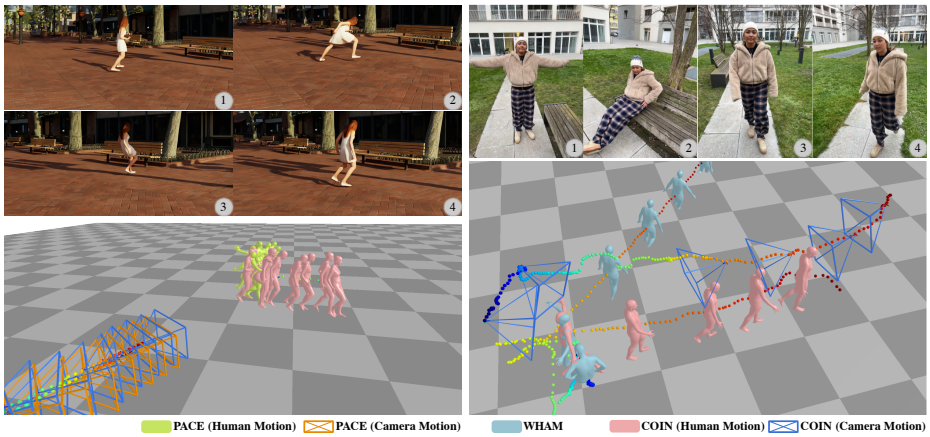

Fig. 3: Qualitative comparisons with state-of-the-art methods. PACE [41] fails to recover a correct trajectory (left). WHAM [81] estimates the wrong walking direction of the person (right). Our approach, COIN, recovers the human and camera motion accurately in both scenarios.

performance because it is not able to accurately estimate the camera trajectories. Noise optimization is better than vanilla SDS but worse than COIN, which generates consistent motion priors. Vanilla SDS is replacing $\mathcal{L}_{\text{COIN}}$ with the SDS loss and keeping the rest of the settings the same. These results demonstrate the effectiveness of COIN over other motion diffusion baselines.

Impact of Dynamic Controlled Sampling. To study the effectiveness of controlled sampling, we compare COIN with and without using the controlled denoiser. Quantitative results are shown in Tabs. 1 and 4. We observe that controlled sampling significantly improves the performance of COIN. Specifically, COIN with controlled sampling reduces the W-MPJPE and ATE-S by **570.5** mm and **167.9** mm, showing **69.3**% and **30.3**% relative improvement, respectively. This demonstrates that controlled sampling is able to generate high-quality motion priors that align with the observed evidence and help improve human and camera motion estimation.

Impact of Soft Inpainting. We further study the effectiveness of soft inpainting. Quantitative comparisons are shown in Tabs. 1 and 4. We observe that while soft inpainting also improves W-MPJPE, it is much more effective than controlled sampling in terms of local body motions. Specifically, COIN with soft inpainting reduces the PA-MPJPE by **4.7** mm.

5 Conclusion

In this paper, we propose COIN, a diffusion-based optimization framework for global human and camera motion estimation from dynamic cameras. We identify the inconsistency problem of distilling knowledge from the diffusion model

with conventional SDS loss. To address this issue, COIN uses a controlled denoiser combined with soft inpainting to distill a high-quality, well-aligned, and consistent motion prior. To further address the scale ambiguity of the camera trajectory, we develop a novel human-scene relation loss that imposes consistency among the human motion, camera motion, and scene features. Comprehensive experiments on challenging synthetic and real-world datasets demonstrate the effectiveness of COIN, which outperforms the SOTA by a large margin in recovering accurate global human motion and camera motion.

Limitations and Future Works: While COIN is able to jointly optimize camera trajectories and global human motions, it requires initialization from SLAM. If the SLAM method fails catastrophically, COIN may not be effective. Additionally, we found COIN fails under severe occlusions where there are several unseen frames and the diffusion model cannot provide consistent guidance. Another limitation is that COIN is an optimization framework, so it is unsuitable for real-time applications. Since the denoising diffusion models have shown their power to model data distributions in many different domains, looking forward we can learn the joint distribution of humans and cameras with the diffusion model to additionally denoise the artifacts in camera motions. Such a joint distribution model also holds the potential of real-time human and camera motion estimation via guided sampling with few DDIM denoising steps.

Appendix

A Controlled Denoiser

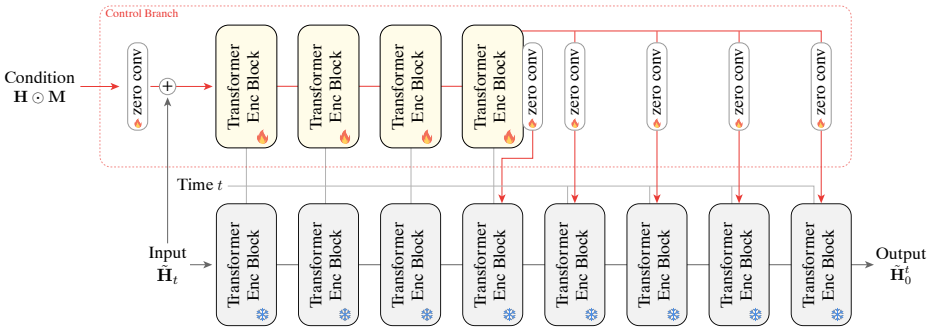

Fig. 4: Architecture of the controlled denoiser.

Architecture Details. The detailed architecture of the proposed controlled denoiser is illustrated in Fig. 4. We adopt a pre-trained transformer-based motion diffusion model as our backbone model. We create a trainable copy of the first 4

encoder blocks. The condition is first encoded by a zero-initialized convolution layer and then concatenates with the input latent motion $\tilde{\mathbf{H}}_t$. The outputs are followed by 5 zero convolution layers and added to the last 5 encoder blocks.

Training Details. We use the AMASS [56] dataset to train the controlled denoiser. To simulate noisy motions in our application, we add Gaussian noise to the conditions. For the root orientation, the noise level is set to 0.05. For the body pose, the noise level is set to 0.01. For the translation, the noise level is set to 0.1. To simulate occlusions, we randomly mask the conditions. With 0.5 probability, all global trajectories are masked out; with 0.5 probability, all global root orientations are masked out. The probabilities of the above two cases are calculated independently, *i.e.*, it is possible to mask out the trajectories and orientations at the same time. With 0.2 probability, the lower half of the body is masked out; with 0.2 probability, the entire local pose is masked out; with 0.5 probability, we randomly mask body joints, and each joint is masked with 0.3 probability.

B Ablation Study

Impact of Each Component. To comprehensively evaluate the impact of each component of COIN, we further conduct ablation studies on the EMDB [37] and HCM [41] datasets. Quantitative results are shown in Tabs. 5 and 7. COIN shows consistent improvement against other diffusion-based baselines.

Table 5: Global human motion estimation on the EMDB dataset.

Method	PA-MPJPE ↓	W-MPJPE$_{100}$ ↓	WA-MPJPE$_{100}$ ↓	RTE ↓	ROE ↓
Noise Optimization	53.9	873.8	275.8	10.4	96.4
Guided Sampling	107.5	1713.9	462.8	7.2	71.5
Vanilla SDS	64.5	1310.3	520.0	12.3	83.0
COIN w/o Controlled Sampling	39.6	815.2	338.7	7.8	44.3
COIN w/o Dynamic Control	36.4	441.2	162.1	4.1	40.2
COIN w/o Soft Inpainting	35.1	495.4	176.8	4.8	43.6
COIN w/o $\mathcal{L}_{\text{HSR}}$	33.0	461.3	162.6	4.0	38.4
COIN	**32.7**	**407.3**	**152.8**	**3.5**	**34.1**

To further evaluate the effect of each individual loss, we report the W-MPJPE on the RICH dataset.

COIN (full)	w/o $\mathcal{L}_{2D}$	w/o $\mathcal{L}_{3D}$	w/o $\mathcal{L}_{\beta}$	w/o $\mathcal{L}_{smooth}$	w/o $\mathcal{L}_{contact}$	w/o $\mathcal{L}_{SDS}$	w/o $\mathcal{L}_{HSR}$
254.5	448.7	329.4	279.9	256.1	270.3	480.6	273.0

Error Distribution. We further present the error distribution on the EMDB dataset to show more details of the COIN predictions. We also plot the error distribution of WHAM [81] for comparison. The scatter plot is shown in Fig. 5. Here we follow the evaluation protocol of EMDB and evaluate W-MPJPE and WA-MPJPE per 100 frames. It is shown that COIN is more robust and has fewer outlier predictions.

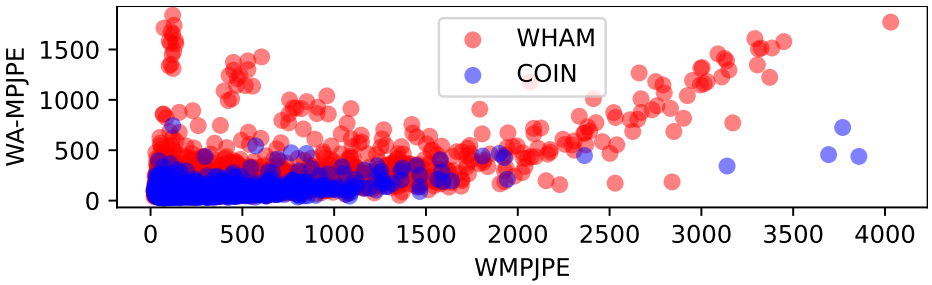

Fig. 5: Error distributions on the EMDB dataset.

SLAM vs SfM. Compared to SLAM, SfM methods are stronger baselines for camera motion estimation. Because our cases always contain dynamic objects, we used ParticleSfM [117] for its ability to handle dynamic objects and compared it to SLAM. Comparisons are shown in Tab. 6. Note that SfM methods run extremely slow compared to SLAM. Given a video with 700 frames on the RICH dataset [25], ParticleSfM takes over 17 hours while DROID-SLAM only needs 4 minutes. For videos on the EMDB dataset with more than 2000 frames, ParticleSfM took a few days to finish, and could not converge for many. Hence, SLAM is a more viable choice for in-the-wild videos. If we replace DROID-SLAM with ParticleSfM, on the converged videos, the baseline results improve by 200 mm. However, it has minimal impact on COIN demonstrating the robustness of our method to SLAM errors. We would like the emphasize that ParticleSfM could not converge on many of the EMDB videos and the results below are only the subset where it converged.

Table 6: SLAM vs. ParticleSfM on the converged subset of the EMDB dataset.

	HybrIK + SLAM	HybrIK + SfM	COIN (SLAM)	COIN (SfM)
W-MPJPE	643.9	439.0	350.1	330.9

Table 7: Global human motion estimation on the HCM dataset.

Method	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	W-RJE ↓	ACCEL ↓
Noise Optimization	66.0	813.9	328.9	794.7	10.2
Guided Sampling	118.2	1653.0	635.4	1626.7	24.7
Vanilla SDS	59.0	1108.2	569.2	1102.6	11.8
COIN w/o Controlled Sampling	47.6	904.8	428.5	898.9	11.0
COIN w/o Dynamic Control	47.4	486.9	239.7	477.5	10.2
COIN w/o Soft Inpainting	48.6	487.1	264.1	478.0	10.8
COIN w/o $\mathcal{L}_{\text{HSR}}$	47.0	488.5	219.3	479.4	10.1
COIN	**45.5**	**479.9**	**212.1**	**470.7**	**10.1**

C Global Optimization

Here we detail our optimization formulation for the reconstruction of global human and camera motion. Simultaneously optimizing both camera motion and global human motion can result in local minima. To address this challenge, we follow PACE [41] and adopt a multi-stage optimization pipeline. Before running optimization, we initialize the global human motion with the noisy observations using the controlled denoiser. We randomly sample a Gaussian noise and run DDPM to generate the global motion. In stage 1, we optimize only the first frame camera parameters (R_0, h_0), camera scale s, and the body shape β. In stage 2, we optimize the first frame camera parameters (R_0, h_0), camera scale s, the body shape β, and the global human motion $\mathbf{H}$. In stage 3, we jointly optimize the full camera trajectory along with the global human motion. Given a long video, we split it into windows of $T = 128$ frames. We use 16 overlapping frames to help reduce discontinuities across windows. The mask $\mathbf{M}$ is defined by thresholding the confidence scores of the detected 2D keypoints. The threshold is 0.3. Each stage is run for 500 steps. The learning rates of the 3 stages are 0.01, 0.01, and 0.001, respectively. We use the Adam solver for optimization. Implementation is in PyTorch.

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR (2015) 3
2. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: ICCV (2019) 4
3. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: CVPR Workshops (2018) 4
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016) 3
5. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: ECCV. pp. 387–404. Springer (2020) 4
6. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: ECCV (2020) 3

7. Choi, H., Moon, G., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: CVPR (2021) 3
8. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: ECCV (2020) 3
9. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020) 10
10. Dabral, R., Shimada, S., Jain, A., Theobalt, C., Golyanik, V.: Gravity-aware monocular 3d human-object reconstruction. In: ICCV (2021) 3
11. Fabbri, M., Lanzi, F., Calderara, S., Alletto, S., Cucchiara, R.: Compressed volumetric heatmaps for multi-person 3d pose estimation. In: CVPR (June 2020) 3
12. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV (2015) 4
13. Gärtner, E., Andriluka, M., Xu, H., Sminchisescu, C.: Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In: CVPR (2022) 3
14. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: ICCV (2023) 12
15. Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., Ororbia, A.G.: A neural temporal model for human motion prediction. In: CVPR (2019) 4
16. Guler, R.A., Kokkinos, I.: HoloPose: Holistic 3d human reconstruction in-the-wild. In: CVPR (2019) 3
17. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human POSEitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: CVPR (2021) 4
18. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion in-betweening. ACM Transactions on Graphics (TOG) **39**(4), 60–1 (2020) 4
19. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: ICCV (2021) 4
20. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: ICCV. pp. 2282–2292 (2019) 4
21. He, C., Saito, J., Zachary, J., Rushmeier, H., Zhou, Y.: Nemf: Neural motion fields for kinematic animation. In: NeurIPS (2022) 2, 4, 11
22. Henning, D.F., Laidlow, T., Leutenegger, S.: Bodyslam: Joint camera localisation, mapping, and human motion tracking. In: ECCV (2022) 4
23. Hernandez, A., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. In: CVPR (2019) 4
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPs (2020) 2
25. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: CVPR (2022) 3, 11, 17
26. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: CVPR (2023) 4
27. Iqbal, U., Molchanov, P., Kautz, J.: Weakly-supervised 3d human pose learning via multi-view images in the wild. In: CVPR (2020) 3
28. Iqbal, U., Xie, K., Guo, Y., Kautz, J., Molchanov, P.: KAMA: 3D keypoint aware body mesh articulation. In: 3DV (2021) 3
29. Isogawa, M., Yuan, Y., O'Toole, M., Kitani, K.M.: Optical non-line-of-sight physics-based 3d human pose estimation. In: CVPR (2020) 3

30. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: CVPR (2016) 4
31. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020) 3
32. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In: 3DV (2021) 3
33. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018) 3
34. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019) 3
35. Karunratanakul, K., Preechakul, K., Aksan, E., Beeler, T., Suwajanakorn, S., Tang, S.: Optimizing diffusion noise can serve as universal motion priors. arXiv preprint arXiv:2312.11994 (2023) 4
36. Kaufmann, M., Aksan, E., Song, J., Pece, F., Ziegler, R., Hilliges, O.: Convolutional autoencoders for human motion infilling. In: 3DV (2020) 4
37. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In: ICCV (2023) 3, 11, 16
38. Khurana, T., Dave, A., Ramanan, D.: Detecting invisible people. In: ICCV. pp. 3174–3184 (2021) 4
39. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: Video inference for human body pose and shape estimation. In: CVPR (2020) 3, 4
40. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: ICCV (2021) 3
41. Kocabas, M., Yuan, Y., Molchanov, P., Guo, Y., Black, M.J., Hilliges, O., Kautz, J., Iqbal, U.: PACE: Human and motion estimation from in-the-wild videos. In: 3DV (2024) 1, 2, 3, 4, 10, 11, 12, 13, 14, 16, 18
42. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019) 3
43. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019) 3
44. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021) 3
45. Kundu, J.N., Rakesh, M., Jampani, V., Venkatesh, R.M., Babu1, R.V.: Appearance consensus driven self-supervised human mesh recovery. In: ECCV (2020) 3
46. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3D and 2D human representations. In: CVPR (2017) 3
47. Li, J., Bian, S., Xu, C., Liu, G., Yu, G., Lu, C.: D &d: Learning human dynamics from dynamic camera. In: ECCV (2022) 2, 4
48. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: CVPR (2021) 3, 5, 8, 12, 13
49. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: ECCV (2022) 3
50. Li, Z., Zhou, Y., Xiao, S., He, C., Huang, Z., Li, H.: Auto-conditioned recurrent networks for extended complex human motion synthesis. arXiv preprint arXiv:1707.05363 (2017) 4

51. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021) 3
52. Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., Tang, S.: 4d human body capture from egocentric video via 3d scene grounding. In: 3DV (2021) 3, 4
53. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. SIGGRAPH Asia **34**(6), 248:1–248:16 (2015) 6
54. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: ACCV (2020) 3
55. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. NeurIPS **34** (2021) 4
56. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: ICCV (2019) 9, 16
57. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018) 4
58. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR (2017) 4
59. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV (2017) 3
60. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: XNect: Real-time multi-person 3D motion capture with a single RGB camera. In: SIGGRAPH (2020) 3
61. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H., Xu, W., Casas, D., Theobalt, C.: VNect: Real-time 3D human pose estimation with a single RGB camera. In: SIGGRAPH (2017) 3
62. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: ICCV (2019) 3
63. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: ECCV (2020) 3
64. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self contact and human pose. In: CVPR (2021) 3
65. Müller, L., Ye, V., Pavlakos, G., Black, M., Kanazawa, A.: Generative proxemics: A prior for 3d social interaction from images. arXiv preprint arXiv:2306.09337 (2023) 4
66. Opensfm - a structure from motion library. https://github.com/mapillary/OpenSfM (2021), https://github.com/mapillary/OpenSfM 2
67. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019) 3
68. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: ICCV (2019) 3
69. Pavlakos, G., Weber, E., Tancik, M., Kanazawa, A.: The one where they reconstructed 3d humans and environments in tv shows. In: ECCV (2022) 4
70. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: CVPR (2018) 3
71. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. In: BMVC (2018) 4

72. Payer, C., Neff, T., Bischof, H., Urschler, M., Stern, D.: Simultaneous multi-person detection and single-person pose estimation with a single heatmap regression network. In: ICCV PoseTrack Workshop (2017) 3
73. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: ICCV (2021) 4
74. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2023) 2, 6
75. Reddy, N.D., Guigues, L., Pischulini, L., Eledath, J., Narasimhan, S.: Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In: CVPR (2021) 3
76. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: ICCV (2021) 2, 3, 4
77. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net: Localization-classification-regression for human pose. In: CVPR (2017) 3
78. Rong, Y., Liu, Z., Li, C., Cao, K., Change Loy, C.: Delving deep into hybrid annotations for 3d human recovery in the wild. In: ICCV (2019) 3
79. Sárándi, I., Hermans, A., Leibe, B.: Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In: WACV (2023) 10
80. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: Physcap: Physically plausible monocular 3d motion capture in real time. In: SIGGRAPH (2020) 3
81. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: CVPR (2024) 1, 2, 3, 4, 11, 12, 13, 14, 17
82. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015) 4
83. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 7
84. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: ECCV (2020) 3
85. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021) 3
86. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In: CVPR (2023) 4, 12
87. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3D people in depth. In: CVPR (2022) 3
88. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., , Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: ICCV (2019) 3
89. Teed, Z., Deng, J.: DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. NeurIPs (2021) 2, 5, 12, 13
90. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. NeurIPS (2023) 2, 12
91. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR 2023 (2022) 2, 4, 8
92. Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H.: Learning to generate long-term future via hierarchical prediction. In: ICML (2017) 4
93. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. In: CVPR (2021) 3
94. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body and hands in the wild. In: CVPR (2019) 3
95. Xie, K., Wang, T., Iqbal, U., Guo, Y., Fidler, S., Shkurti, F.: Physics-based human motion estimation and synthesis from videos. In: ICCV (2021) 3

96. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. In: ICLR (2024) 4
97. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: ICCV (2019) 3
98. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: ECCV (2018) 4
99. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: CVPR (2023) 2, 4, 10, 12, 13
100. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: CVPR (2022) 2, 4, 11, 12, 13
101. Yuan, Y., Kitani, K.: Diverse trajectory forecasting with determinantal point processes. In: ICLR 2020 (2019) 4
102. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: ECCV (2020) 4
103. Yuan, Y., Kitani, K.: Residual force control for agile human behavior imitation and extended motion synthesis. In: NeurIPS (2020) 4
104. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: ICCV (2023) 4
105. Yuan, Y., Wei, S.E., Simon, T., Kitani, K., Saragih, J.: Simpoe: Simulated character control for 3d human pose estimation. In: CVPR (2021) 3
106. Zanfir, A., Bazavan, E.G., Xu, H., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In: ECCV (2020) 3
107. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes the importance of multiple scene constraints. In: CVPR (2018) 3
108. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3d sensing of multiple people in natural images. In: NeurIPS (2018) 3
109. Zanfir, M., Zanfir, A., Bazavan, E.G., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Thundr: Transformer-based 3d human reconstruction with markers. In: ICCV (2021) 3
110. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: ICCV (2021) 3
111. Zhang, J., Yu, D., Liew, J.H., Nie, X., Feng, J.: Body meshes as points. In: CVPR (2021) 3
112. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023) 8
113. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 4
114. Zhang, S., Bhatnagar, B.L., Xu, Y., Winkler, A., Kadlecek, P., Tang, S., Bogo, F.: Rohm: Robust human motion reconstruction via diffusion. In: CVPR (2024) 4
115. Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: ICCV (2021) 4
116. Zhang, T., Huang, B., Wang, Y.: Object-occluded human shape and pose estimation from a single color image. In: CVPR (2020) 3

117. Zhao, W., Liu, S., Guo, H., Wang, W., Liu, Y.J.: Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In: European conference on computer vision (ECCV) (2022) 17
118. Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., Zhou, X.: SMAP: Single-shot multi-person absolute 3d pose estimation. In: ECCV (2020) 3
119. Zhou, Y., Habermann, M., Habibie, I., Tewari, A., Theobalt, C., Xu, F.: Monocular real-time full body capture with inter-part correlations. In: CVPR (2021) 3