Masked Modeling for Human Motion Recovery Under Occlusions

Zhiyin Qian[1*] Siwei Zhang[2†] Bharat Lal Bhatnagar[2] Federica Bogo[2] Siyu Tang[1]
[1]ETH Zürich [2]Meta Reality Labs

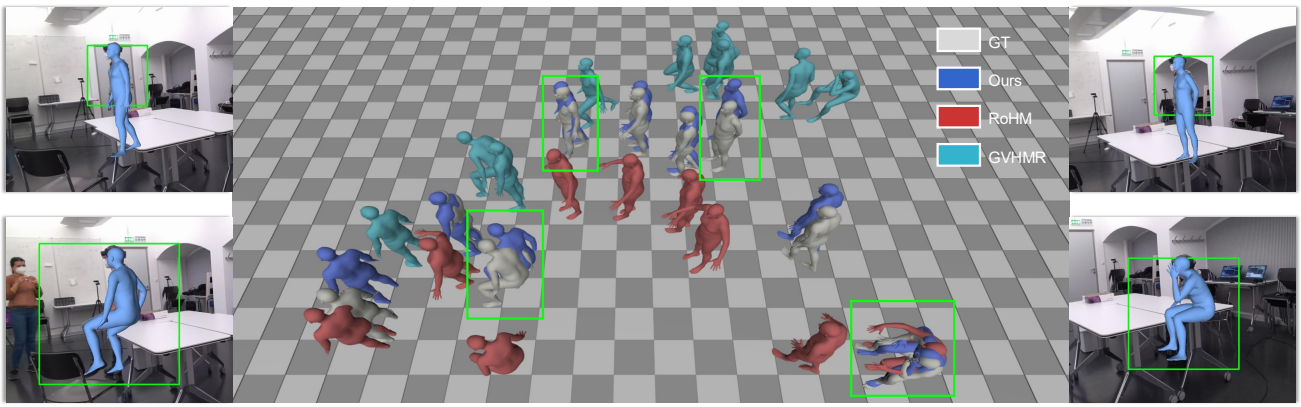

Figure 1. **Overview.** Given a monocular video captured from a static camera, MoRo robustly reconstructs accurate and physically plausible human motion (middle), even under challenging occlusion scenarios where detections cover only portions of the subject (on both sides). Leveraging masked modeling, our method iteratively synthesizes motion in a globally consistent coordinate by integrating both visual cues and motion priors, facilitated by our cross-modality training strategy. Compared to baselines such as GVHMR [57] and RoHM [79], MoRo consistently achieves better performance in terms of both per-frame accuracy and global motion realism, particularly in the presence of occlusions.

Abstract

Human motion reconstruction from monocular videos is a fundamental challenge in computer vision, with broad applications in AR/VR, robotics, and digital content creation, but remains challenging under frequent occlusions in real-world settings. Existing regression-based methods are efficient but fragile to missing observations, while optimization- and diffusion-based approaches improve robustness at the cost of slow inference speed and heavy preprocessing steps. To address these limitations, we leverage recent advances in generative masked modeling and present ***MoRo****: **M**asked m**O**deling for human motion **R**ecovery under **O**cclusions. MoRo is an occlusion-robust, end-to-end generative framework that formulates motion reconstruction as a video-conditioned task, and efficiently recover human motion in a consistent global coordinate system from RGB videos. By masked modeling, MoRo naturally handles occlusions while enabling efficient, end-to-end inference. To overcome the scarcity of paired video–motion data, we design a cross-modality learning scheme that learns multi-modal priors from a set of heterogeneous datasets: (i) a trajectory-aware motion prior trained on MoCap datasets, (ii) an image-conditioned pose prior trained on image-pose datasets, capturing diverse frame-level poses, and (iii) a video-conditioned masked transformer that fuses motion and pose priors, finetuned on video–motion datasets to integrate visual cues with motion dynamics for robust inference. Extensive experiments on EgoBody and RICH demonstrate that MoRo substantially outperforms state-of-the-art methods in accuracy and motion realism under occlusions, while performing on-par in non-occluded scenarios. MoRo achieves real-time inference at 70 FPS on a single H200 GPU. Project page: https://mikeqzy.github.io/MoRo.*

1. Introduction

Reconstructing 3D human pose and motion from monocular RGB inputs is a long-standing problem in computer

[*]All data access, experiments, and model training were conducted at ETH Zürich.

[†]This work was completed while SZ was a postdoctoral researcher at ETH Zürich.

vision [3, 8–11, 15, 24, 26–29, 35, 36, 38, 41, 46, 47, 50, 54, 69, 74, 78, 80, 82], with broad applications in augmented and virtual reality, assistive robotics, and healthcare. However, the limited field of view of monocular cameras often leads to body occlusions when capturing people moving in real-world environments, making motion reconstruction challenging. Despite the recent rapid progress in this area driven by advances in deep neural network architectures [65], existing methods still struggle with such occlusions.

Most regression-based approaches [10, 25, 27, 47] offer end-to-end fast inference, but perform poorly under heavy occlusions. Recent methods tackle dynamic camera scenarios [30, 57, 59, 66, 71, 73] by jointly estimating human and camera motion in global space, but without explicitly addressing occlusion scenarios. Generative modeling is well-suited to tackle the motion ambiguities caused by occlusions. Optimization-based methods such as HuMoR [53] and PhaseMP [58] incorporate VAE-based motion priors within optimization loops [53, 58], yielding more robust performance than regressors but at the cost of slow inference, sensitivity to initialization, and susceptibility to local minima. RoHM [79] surpasses optimization-based methods in speed and robustness by casting the task as conditional diffusion, but does not provide real-time performance, still fails under severe occlusions, and relies on pose initialization and precomputed body visibility. Moreover, most of these methods depend solely on precomputed 2D/3D joints and discard the rich visual context available in videos. These limitations underscore the need for occlusion-robust, end-to-end models capable of real-time inference.

To fill this gap we propose **MoRo**, a generative framework for robust, efficient motion reconstruction from videos, which builds on recent advances in Masked Generative Transformers [5, 12, 19, 52, 75, 83]. Namely, MoRo reformulates motion reconstruction as a video-conditioned generative task via masked modeling. Masked modeling with transformers has been widely adopted in text [12], image [5], and motion [19, 52, 83] generation. By randomly masking sequence segments, the model learns to reconstruct missing parts - an intuitive fit for handling occlusions. Unlike optimization- or diffusion-based methods [53, 58, 79], which are slow and initialization-sensitive, masked modeling can enable efficient, end-to-end inference. While prior works such as GenHMR [55] and MEGA [17] apply this paradigm to single-frame mesh recovery, extending it to video is far more challenging: it requires not only resolving per-frame ambiguities but also modeling long-term dynamics across local and global pose spaces while remaining faithful to visual evidence under severe occlusions.

Directly learning the video-to-motion mapping under body occlusions as in [16] is challenging due to the scarcity of paired video–motion data. To address this, we decompose the learning process across diverse modalities spanning motion, image, and video datasets, and integrate them into a unified framework with end-to-end inference. Following [16], we represent 3D human meshes by discrete local pose tokens using a pre-trained Vector Quantized Variational Autoencoder (VQ-VAE) [64]. To recover motion from missing observations, it is crucial to model natural human dynamics. We begin by training a trajectory-aware motion prior on large-scale MoCap datasets [44] with masked modeling, where the model jointly denoises a noisy input root trajectory and predicts missing local pose tokens. To overcome the limited pose diversity in MoCap, we then train an image-conditioned pose prior on large-scale image–pose datasets [1, 22, 40, 45] for pose reconstruction, while the image encoder of this prior also estimates a coarse global trajectory that serves as input to the motion prior. Finally, we fine-tune a video-conditioned masked motion transformer — combining the pretrained motion prior, pretrained image-conditioned pose prior, and a cross-modality decoder — on video datasets [2, 77] via masked modeling, enabling the recovery of missing pose tokens and denoising of the global trajectory conditioned on video evidence. A multi-step inference process iteratively recovers pose tokens from video evidence while refining the global trajectory. Unlike prior occlusion-handling methods [53, 58, 76, 79] that overlook visual context in motion prior learning, MoRo unifies learning across diverse datasets and modalities in a single end-to-end framework, eliminating reliance on preprocessing and efficiently leveraging multi-modality priors to enhance robustness for motion recovery under occlusions.

In summary, our contributions are: 1) MoRo, a novel generative framework that leverages masked modeling for robust and efficient motion recovery from monocular videos; 2) a cross-modality learning scheme that fuses multi-modal priors learnt across motion, image and video data, effectively learning a video-conditioned motion distribution. Extensive evaluations show that MoRo significantly outperforms state-of-the-art methods in both reconstruction accuracy and motion realism in challenging occlusion cases, while achieving comparable performance in non-occluded scenarios.

2. Related Work

Human mesh recovery (HMR) from a single image has seen significant progress in recent years. We can distinguish regression-based methods [8, 9, 18, 24, 28, 29, 32, 33, 35, 38, 39, 46, 48, 56, 69, 74, 82], optimization-based methods [3, 15, 34, 50] and hybrid methods [31, 60]. Most methods regress SMPL [42] or SMPL-X [50] parameters, while others predict non-parametric mesh vertices [8, 9, 39, 46] or arbitrary human volume points [56] from images. Recently, VQ-HPS [16] and TokenHMR [14] reformulate HMR from continuous regression to discrete classification by tokenizing human poses, showing improved accuracy. HMR methods vary in focus: most aim for higher accuracy in generic sce-

narios, some enhance camera modeling [29, 38, 49], and others handle occlusions and truncations [28, 33, 67]. Building on tokenized pose representations, MEGA [17] and GenHMR [55] employ generative masked modeling to resolve pose ambiguities, producing multiple hypotheses from a single image. However, these approaches remain limited to static images and cannot model temporal correlations.

Human motion reconstruction from videos aims at estimating plausible 3D human motion from frames. Dealing with occlusions in the temporal domain is even more challenging. Early regression-based methods [7, 10, 25, 43, 47, 68, 72] primarily predict local motion in the camera space without modeling the global trajectory, thus exhibiting motion artifacts. Other optimization-based methods [53, 58, 76] refine noisy per-frame estimates using motion priors and/or scene constraints, improving robustness under occlusions. However, they are slow, sensitive to local minima, and require extensive manual tuning. Moreover, their reliance on noisy per-frame estimates makes them fragile when the initialization is unreliable. More recently, diffusion-based approaches such as RoHM [79] tackle motion reconstruction under occlusions by conditioning on partial observations, yet they still rely on per-frame initialization and remain too slow for real-time use. In contrast, we propose to leverage the generative masked modeling framework to enable end-to-end and real-time inference. Another recent line of work addresses dynamic camera scenarios. Some train regressors [57, 59], while others integrate motion priors with SLAM-based reconstructions in optimization frameworks [30, 37, 71] to jointly estimate human and camera trajectories. However, when applied to videos with occlusions even under static cameras, these methods struggle to robustly reconstruct consistent motion (as shown in Sec. 5.5).

Generative masked modeling. Masked modeling, initially introduced in BERT [12] for language tasks, was later adapted to vision through masked autoencoders [20], where models learn to reconstruct masked tokens from visible context. Building on this idea, masked generative modeling extends the paradigm by starting from a fully masked sequence and progressively generating tokens in fixed steps [5, 6]. It has been applied to human motion generation [19, 23, 51, 52], achieving state-of-the-art performance while being significantly faster than diffusion-based methods. Recent works like GenHMR [55] and MEGA [17] extend the idea to human mesh recovery to generate multiple pose hypotheses from a single image, demonstrating its effectiveness when dealing with amgituities. Still, these methods are limited to static images. In contrast, we further extend the generative masked modeling framework to the video domain, reconstructing natural human motions from videos under occlusions.

3. Motion Representation

SMPL-X [50] is a parametric body model that represents the 3D human body as a function $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\theta}_h, \boldsymbol{\phi})$, which returns a triangle mesh $\mathcal{M}$ with 10,475 vertices. It is parameterized by global translation $\boldsymbol{\gamma}$, global orientation $\boldsymbol{\Phi}$, body pose $\boldsymbol{\theta}$, body shape $\boldsymbol{\beta}$, hand pose $\boldsymbol{\theta_h}$ and facial expression $\boldsymbol{\phi}$. In this paper we consider only the main body parts while omitting $\boldsymbol{\theta}_h$ and $\boldsymbol{\phi}$.

3D Body Mesh Tokenization. Following prior works [14, 16, 17], we utilize a tokenized representation of the human mesh. A local pose tokenizer is pre-trained to learn a discrete latent representation for the human mesh, adopting the convolutional autoencoder architecture from Mesh-VQ-VAE [16]. Given a SMPL-X mesh $\boldsymbol{v} \in \mathbb{R}^{10475\times 3}$ in local coordinates (setting the global orientation and translation to zero to disentangle global trajectory and local pose), the pose tokenizer encoder maps it into latent embeddings $\boldsymbol{z} \in \mathbb{R}^{P\times L}$, where $P = 87$ is the number of tokens and $L = 9$ is the dimension of each token. Each latent embedding $\boldsymbol{z}_i$ is then quantized into a discrete token $\tilde{\boldsymbol{z}}_i$ by finding its nearest neighbor in the codebook $\mathcal{C}$ of size 512, as $\tilde{\boldsymbol{z}}_i = \arg\min_{\boldsymbol{c}_k\in\mathcal{C}} \|\tilde{\boldsymbol{z}}_i - \boldsymbol{c}_k\|_2^2$. The quantized tokens $\tilde{\boldsymbol{z}}$ are mapped back to a human mesh $\tilde{\boldsymbol{V}}$ by a symmetric decoder. The local pose tokenizer is trained on AMASS [44], BEDLAM [2] and MOYO [63], providing a strong prior on plausible human meshes.

Motion representation. We represent a motion sequence of T frames by $\boldsymbol{X} = (\boldsymbol{R}, \boldsymbol{Z})$, where $\boldsymbol{R} \in \mathbb{R}^{T\times 9}$ and $\boldsymbol{Z} \in \mathbb{R}^{T\times P\times L}$ denote the pelvis global trajectory, and quantized local body tokens, respectively. For frame t, the global trajectory $\boldsymbol{R}^t$ consists of the SMPL-X global orientation $\boldsymbol{\Phi} \in \mathbb{R}^6$ in 6D representation [81] and the translation $\boldsymbol{\gamma} \in \mathbb{R}^3$. The tokenized local body pose $\boldsymbol{Z}^t \in \mathbb{R}^{P\times L}$ is obtained from the pose tokenizer, consisting of P discrete pose tokens with the dimension of each token as L.

4. Method

We introduce MoRo, a novel generative masked modeling framework for 3D human motion recovery from monocular videos under body occlusions. Given a monocular video $\boldsymbol{I}$ with F frames captured by a static camera, MoRo aims to learn a conditional distribution of the human motion $p(\boldsymbol{X}|\boldsymbol{I})$.

MoRo features three main components: an image encoder pretrained on large-scale image-pose datasets for visual conditioning and image-conditioned pose prior learning (Sec. 4.1), a motion encoder pretrained on large-scale MoCap datasets via masked modeling for motion prior learning (Sec. 4.2), and a cross-modal decoder finetuned on video-motion data to fuse cross-modality information and efficiently recovering the human motion from videos (Sec. 4.3). A multi-step inference schedule iteratively predicts pose tokens and refines the global trajectory (Sec. 4.4), further improving motion realism under occlusions. The multi-stage

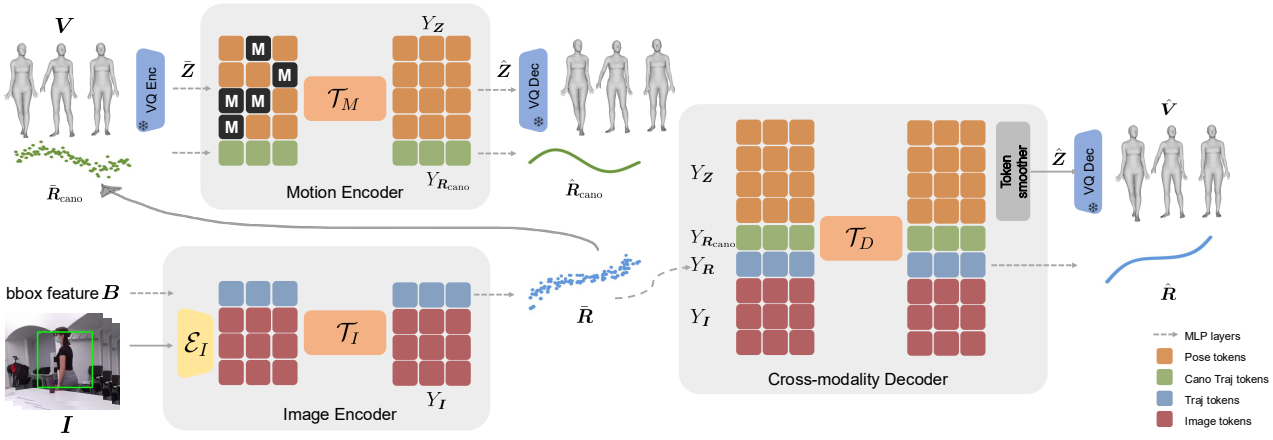

Figure 2. **Overview of our masked transformer,** which consists of three main components: the image encoder, the motion encoder and the decoder. Given a monocular video sequence, we utilize the image encoder to extract per-frame image features and estimate a coarse global trajectory, which is canonicalized and serves as the input to the motion encoder (Sec. 4.1). Along with masked local pose tokens, the motion encoder encodes a trajectory-aware motion prior via recovering the complete local pose tokens and denoising the global trajectory (Sec. 4.2). The cross-modality decoder fuses the intermediate feature from both encoders via a spatial-temporal transformer to refine the camera-space global trajectory and predict a conditional categorical distribution for sampling the local pose tokens, which are then smoothed for enhanced motion realism (Sec. 4.3).

training scheme is explained in Sec. 4.5. Fig. 2 shows an overview of the proposed approach.

4.1. Per-frame Image Conditioning

The image encoder processes each frame to extract image features and estimate a coarse global trajectory. A ViT-H/16 [13] backbone $\mathcal{E}_I$ initialized with pretrained weights from ViTPose [70] encodes the cropped video frames $\boldsymbol{I}$ into image tokens $\mathcal{E}_I(\boldsymbol{I}) \in \mathbb{R}^{F\times 192\times 1280}$, where 192 denotes the number of image tokens. Inspired by [38, 78], the additional bounding box feature $\boldsymbol{B} = (\boldsymbol{b}_x, \boldsymbol{b}_y, \boldsymbol{s})/\boldsymbol{f} \in \mathbb{R}^{F\times 3}$ is utilized to better infer the global positional information, where $\boldsymbol{b}_x, \boldsymbol{b}_y$ denote the bounding box center x-y coordinates relative to the principle point, $\boldsymbol{s}$ is the bounding box size in the original full image, and $\boldsymbol{f}$ is the focal length.

Tokens from $\mathcal{E}_I(\boldsymbol{I})$ and $\boldsymbol{B}$ are projected to the same latent dimension of 1024 by linear layers, concatenated and fed into a transformer network $\mathcal{T}_I$ to further model the visual context:

$$Y_{\boldsymbol{I}}, \bar{\boldsymbol{R}} = \mathcal{T}_I(\mathcal{E}_I(\boldsymbol{I}), \boldsymbol{B}), \tag{1}$$

where $Y_{\boldsymbol{I}} \in \mathbb{R}^{F\times 192\times 1024}$ denotes the encoded latent image feature, serving as the visual conditioning for the cross-modal decoder.

The weak-perspective camera parameters $\bar{\boldsymbol{R}}_{\text{crop}}$ in the cropped view are obtained by applying a linear layer to the latent feature at the bounding box token position. These parameters are then converted back to the full camera space using the CLIFF [38] transformation and temporally stacked to form the coarse global trajectory initialization $\bar{\boldsymbol{R}}$. Note that in the image encoder, each frame is processed separately without incorporating temporal information.

Specifically, the image encoder is combined with a pose decoder (architecturally identical to the cross-modal decoder in Sec. 4.3 without modeling temporal information) and pre-trained on image–pose datasets [1, 22, 40, 45] to learn an image-conditioned pose prior via body pose reconstruction. This prior captures diverse poses present in image datasets but absent in video datasets, thereby facilitating video-conditioned motion learning in later stages.

4.2. Trajectory-aware Motion Prior

Directly recovering global human motion from image features leads to degraded motion quality under occlusions (see Sec. 5.6). To address this, we introduce a data-driven motion prior learned from the AMASS motion capture dataset [44], which models natural human dynamics from partial observations, improving temporal consistency and robustness to occlusions while reducing reliance on large-scale paired video–motion data. Provided the noisy global trajectory $\bar{\boldsymbol{R}}$ predicted by the image encoder, inspired by the insight that human local pose is strongly correlated with its movement in the global space, we design the motion prior in a trajectory-aware manner to further model the strong inter-dependencies between local pose and global trajectory.

We transform $\bar{\boldsymbol{R}}$ into a canonicalized coordinate system in which each frame is represented by its motion toward the next frame, yielding the canonicalized global trajectory $\bar{\boldsymbol{R}}_{\text{cano}}$. For each frame t, the canonicalized global orientation

$\bar{\mathbf{\Phi}}^t_{\text{cano}}$ and translation $\bar{\gamma}^t_{\text{cano}}$ are computed as:

$$\begin{aligned}\bar{\mathbf{\Phi}}^t_{\text{cano}} &= \left(\bar{\mathbf{\Phi}}^t\right)^{-1} \bar{\mathbf{\Phi}}^{t+1}, \\ \bar{\gamma}^t_{\text{cano}} &= \left(\bar{\mathbf{\Phi}}^t\right)^{-1} (\bar{\gamma}^{t+1} - \bar{\gamma}^t),\end{aligned} \tag{2}$$

This canonicalization makes the global trajectory independent of the coordinate system and sequence length, which is crucial for motion modeling. We further apply a binary mask $\boldsymbol{M} \in \{0,1\}^{F\times P}$ to partially mask local pose tokens $\boldsymbol{Z}$ following the paradigm of masked modeling, resulting in the corrupted motion $(\bar{\boldsymbol{R}}_{\text{cano}}, \bar{\boldsymbol{Z}})$ where $\bar{\boldsymbol{Z}} = \boldsymbol{M} \odot \boldsymbol{Z}$. For motion pretraining on AMASS, $\bar{\boldsymbol{R}}_{\text{cano}}$ is obtained by manually corrupting the clean global trajectory by adding Gaussian noise to the body orientation and translation (see Supp. Mat. for details).

A motion transformer $\mathcal{T}_M$ aims to recover the complete pose tokens and denoise the global trajectory simultaneously:

$$\hat{\boldsymbol{Z}}, \hat{\boldsymbol{R}}_{\text{cano}}, Y_{\boldsymbol{Z}}, Y_{\boldsymbol{R}_{\text{cano}}} = \mathcal{T}_M(\bar{\boldsymbol{Z}}, \bar{\boldsymbol{R}}_{\text{cano}}), \tag{3}$$

where $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{R}}_{\text{cano}}$ denote the reconstructed pose tokens and trajectory, respectively. $Y_{\boldsymbol{Z}}$ and $Y_{\boldsymbol{R}_{\text{cano}}}$ are their corresponding latent features encoded by $\mathcal{T}_M$, and later fed to the cross-modal decoder (Sec. 4.3) for video-conditioned motion recovery. During motion pretraining, the recovered local pose tokens $\hat{\boldsymbol{Z}}$ are further mapped back to the original SMPL-X mesh sequence by the pose tokenizer for self-supervision in the vertex space.

Unlike previous works [53, 58, 79] where motion priors solely model motion itself, our proposed prior also acts as the motion encoder in the video-conditioned masked transformer and can be fine-tuned with video data, providing strong knowledge of natural human dynamics while conditioning on video inputs.

4.3. Video-conditioned Masked Transformer

Given the per-frame image features, the pre-trained motion prior and image-conditioned pose prior, the video-conditioned masked transformer further incorporates a spatial–temporal transformer $\mathcal{T}_{\mathcal{D}}$ as the cross-modal decoder to fuse multi-modality information from both image and motion encoders to recover global motion. Following the same masked modeling strategy for training the motion prior, it predicts local pose tokens and the clean global body trajectory conditioning on visual observations.

Firstly, the image encoder predicts per-frame image features $Y_{\boldsymbol{I}}$ and the coarse global trajectory $\bar{\boldsymbol{R}}$ (Eq. 1). The canonicalized trajectory $\bar{\boldsymbol{R}}_{\text{cano}}$ obtained from $\bar{\boldsymbol{R}}$, together with partially masked pose tokens, are then processed by the motion encoder to produce the pose features $Y_{\boldsymbol{Z}}$ and trajectory features $Y_{\boldsymbol{R}_{\text{cano}}}$ (Eq. 3). Finally, transformer $\mathcal{T}_D$ models the spatial-temporal correlations among the multi-modal features from image, pose and trajectory, and generate the complete pose token sequence $\hat{\boldsymbol{Z}}$ and a refined global trajectory $\hat{\boldsymbol{R}}$:

$$\hat{\boldsymbol{Z}}, \hat{\boldsymbol{R}} = \mathcal{T}_{\mathcal{D}}(Y_{\boldsymbol{Z}}, Y_{\boldsymbol{R}_{\text{cano}}}, Y_I, Y_{\boldsymbol{R}}), \tag{4}$$

where $Y_{\boldsymbol{R}}$ is obtained by encoding the estimated global trajectory $\bar{\boldsymbol{R}}$ from Eq. 1 by a linear layer.

The predicted pose tokens $\hat{\boldsymbol{Z}}$ are decoded by the pose tokenizer into the original SMPL-X mesh space, and then combined with the reconstructed global trajectory $\hat{\boldsymbol{R}}$ to produce the final reconstructed motion in the global space.

Pose token smoother network. Due to the discretization during tokenization, the generated motion derived from per-frame pose tokens still exhibits some jittering artifacts. To address this, we inject a learnable smoother network $\mathcal{F}_{\text{smoother}}$ to map the discrete latent representation $\hat{\boldsymbol{Z}}$ picked from the codebook into a continuous representation, before decoding it into the canonical mesh. $\mathcal{F}_{\text{smoother}}$ is implemented as a 2-layer MLP, efficiently alleviating the jittering artifacts (Sec. 5.6).

Architectures. The image encoder $\mathcal{T}_I$ employs a transformer encoder structure following [65]. Both the motion encoder $\mathcal{T}_M$ and the cross-modal decoder $\mathcal{T}_D$ adopt the spatial-temporal transformer architecture DSTFormer from [83]. In addition, we leverage Rotary Positional Embedding (RoPE) [61] by computing the temporal attention based on relative temporal positions, enabling MoRo to handle sequences with variable length during inference.

4.4. Inference

At inference time, the model iteratively recovers masked tokens based on uncertainty of each predicted pose token. The image encoder is first executed once to extract image features and predict a coarse global trajectory $\bar{\boldsymbol{R}}$ by Eq. 1. In the first inference iteration, fully masked pose tokens together with canonicalized $\bar{\boldsymbol{R}}_{\text{cano}}$ are fed into the motion encoder $\mathcal{T}_M$ and decoder $\mathcal{T}_D$ to generate the complete pose tokens and a refined global trajectory $\hat{\boldsymbol{R}}$. We then retain the top-K pose tokens with the highest confidence, along with the refined global trajectory, as input for the next iteration, while the remaining tokens are re-masked for regeneration. For each pose token, the confidence refers to the predicted logits after the softmax layer. The refined global trajectory $\hat{\boldsymbol{R}}$ at each iteration will be canonicalized and fed to the motion encoder to be refined in the next iteration. The process repeats until all tokens are recovered. We adopt $T = 5$ as the number of inference iteration steps. The final pose tokens are decoded into the SMPL-X mesh and transformed to global coordinates using predicted trajectory from the last iteration.

4.5. Multi-stage Training

In order to strike the balance between faithfully recovering motion from available image evidence and generating realistic motions for occluded body parts, the proposed model

is trained in a progressive manner, spanning datasets with different annotation modalities.

Motion Pretraining. The motion encoder is pretrained on AMASS [44]. In addition to random masking adopted by previous works [17, 55], we introduce spatial and temporal masking strategies that either mask all tokens in selected frames or mask specific tokens across all frames during training, better reflecting real-world scenarios where occlusions are typically continuous in space and time.

Image Pretraining. We pretrain the image encoder and the cross-modal decoder on standard image datasets - including Human3.6M [22], MPI-INF-3DHP [45], COCO [40] and MPII [1] - to improve generalization to diverse body poses, which are less represented in video datasets. During image pretraining, motion-related features ($Y_{\boldsymbol{Z}}, Y_{\boldsymbol{R}_{\text{cano}}}$) are simply masked out in the decoder.

Video Fine-tuning. Finally, the full model is fine-tuned on two monocular video-motion datasets, EgoBody [77] and BEDLAM [2]. The spatial-temporal cross-modal decoder leverages the pre-trained motion prior information and image-pose prior information to further model the correlations among multiple modalities from visual inputs, global trajectory and local motion.

Confidence-guided masking. Alongside random masking, we adopt a confidence-guided masking strategy during video fine-tuning. Starting with fully masked inputs, we perform one inference step, re-mask a subset of tokens according to their predicted confidence as in the iterative inference (Sec. 4.4), and run a subsequent prediction round. This reduces the train–test gap and enables the model to recover tokens from imperfect inputs in multi-step inference.

Training objective. MoRo is trained with the cross entropy loss $\mathcal{L}_{\text{ce}}$ for the local pose token classification, local 3D mesh vertex loss $\mathcal{L}_{V_{3D}}$, global trajectory loss $\mathcal{L}_{\text{traj}}$, global 3D joint position loss $\mathcal{L}_{J_{3D}}$ and velocity loss $\mathcal{L}_{j_{3D}}$, 2D keypoint reprojection loss $\mathcal{L}_{J_{2D}}$ and foot skating loss $\mathcal{L}_{\text{fs}}$:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{V_{3D}} + \mathcal{L}_{\text{traj}} + \mathcal{L}_{J_{3D}} + \mathcal{L}_{j_{3D}} + \mathcal{L}_{J_{2D}} + \mathcal{L}_{\text{fs}}, \quad (5)$$

where the global joint losses $\mathcal{L}_{J_{3D}}, \mathcal{L}_{j_{3D}}$ are computed from multiple global trajectory predictions $\bar{\boldsymbol{R}}, \hat{\boldsymbol{R}}_{\text{cano}}, \hat{\boldsymbol{R}}$ from our model and $\mathcal{L}_{\text{fs}}$ penalizes the foot velocity if it exceeds a certain threshold and is in contact with ground to reduce foot skating artifacts. Each loss term is weighted by corresponding weights and applied only when it is applicable in the pretraining stages. Please see Supp. Mat. for more details.

5. Experiments

5.1. Datasets

MoRo is trained on datasets across multiple modalities as described in Sec. 4.5, and evaluated on **EgoBody** [77] and **RICH** [21]. Both EgoBody and RICH capture human motions interacting with various 3D environments, recording multi-modal data streams including third-person view RGB videos, with the human motion annotated with SMPL-X parameters. EgoBody includes a notable amount of occlusion scenarios, and we evaluate the proposed method on a subset of EgoBody third-view videos with severe body occlusions following [79], excluding sequences in the EgoBody training split. This curated subset is denoted as *EgoBody-Occ* and consists of 17 video sequences with a total of 23055 frames. RICH, on the other hand, has relatively uncluttered scenes, resulting in few occlusions in videos. It has been a standard evaluation dataset for evaluating video-based motion reconstruction [30, 57, 59, 66]. We evaluate on 191 test sequences to assess model performance in non-occluded scenarios.

5.2. Implementation Details

The predicted SMPL-X mesh vertices from our method are fitted to SMPL-X parameters using a fast fitting algorithm [62] for evaluation. We use the same bounding box detections across all methods for fair comparison. On EgoBody, bounding boxes are derived from OpenPose 2D keypoints [4], excluding keypoints with confidence below 0.2. On RICH, we adopt the preprocessed bounding boxes provided by [57]. Ground truth camera intrinsics is employed in all methods. We perform extreme cropping augmentation [14] by randomly cropping human body parts from images to further improve the model's robustness to truncated bounding boxes.

5.3. Evaluation Metrics

Accuracy. The local pose and shape accuracy is evaluated via Mean Per Joint Position Error (*MPJPE* in mm), Procrustes-aligned MPJPE (*PA-MPJPE* in mm), and Per Vertex Error (*PVE* in mm). Following [79], we report MPJPE for full-body (*-all*), visible joints (*-vis*) and occluded (*-occ*) joints separately on EgoBody-occ. For global-space reconstruction, prior works [57, 59, 66, 67, 73] often report World MPJPE, which aligns predicted and ground-truth motions by the first frame over each 100-frame segments, thereby underestimating global translation errors in long sequences. Instead, we report Global MPJPE (*GMPJPE* in mm) and Root Translation Error (*RTE* in %) [57] to evaluate long-term global accuracy.

Motion Quality. We report metrics on motion smoothness and foot sliding of the reconstructed motion to evaluate the motion plausibility. Consistent with prior works [57, 59, 66, 67, 73], we report the local acceleration error (Accel, in m/s^2), motion jitter (in m/s^3), and foot sliding (in mm). Additionally, we find that global acceleration error (G-Accel, in m/s^2) better reflects the motion smoothness globally.

5.4. Baselines

We compare MoRo against (1) per-frame pose estimation methods: MEGA [17], TokenHMR [14], PromptHMR [67]†,

and (2) video-based motion reconstruction methods: RoHM [79], WHAM [59], GVHMR [57]. RoHM is a diffusion-based method that relies on per-frame 3D body pose initializations and precomputed occlusion masks. WHAM and GVHMR, though designed for dynamic-camera settings, also applies to static cameras by fixing camera extrinsics to identity. Both output a camera-space trajectory and a world-grounded trajectory, which should differ only by a rigid transform under the static camera setup; however, we find them inconsistent in practice due to an additional refinement step on the world-grounded trajectory - the world-grounded trajectory achieves better motion realism but degrades video–motion alignment (Fig. 3, row 3), while the camera-space trajectory aligns better with video but yields lower motion quality. They evaluate per-frame metrics (PA-MPJPE, MPJPE, PVE) using camera-space predictions and global metrics (RTE, Jitter, Foot-Sliding) using world-grounded predictions, whereas a single model should ideally produce unified outputs consistent across both camera and global space. We therefore report results separately for each prediction, denoted by *-Cam* and *-World*. Please refer to Supp. Mat. for more details.

5.5. Results

Performance on videos with occlusions. Tab. 1 shows the quantitative results on EgoBody-Occ. Our method consistently surpasses baselines in both accuracy and motion quality, demonstrating strong robustness under occlusions.

For local joint accuracy, our model outperforms all baselines on both visible and occluded joints, achieving a **16/31%** MPJPE improvement over the best baseline PromptHMR for visible/occluded body parts, respectively. Although PromptHMR is relatively robust to various bounding box sizes by encoding full-image context and using the bounding box only as a spatial prompt, our method still surpasses it. In terms of global reconstruction, our model exceeds the best baseline RoHM by a large margin, achieving **58%** better global joint reconstruction (GMPJPE). RoHM suffers from poor local pose accuracy (with a high PA-MPJPE) since it ignores visual input during the motion prior learning, whereas our method effectively addresses visual evidence and the motion prior within one unified framework.

Regarding motion realism, our method produces remarkably plausible motion with the least jitters and the second least foot sliding. RoHM generates fixed-length clips, leading to temporal discontinuities for long sequences, while our RoPE-based [61] transformer maintains consistency over arbitrary-length videos by attending to local 60-frame contexts. As mentioned in Sec. 5.4, WHAM and GVHMR produce two sets of inconsistent outputs: camera-space predictions (-Cam) yield better per-frame accuracy but suffer from motion jitter, while world-grounded predictions (-World) improve motion realism by additional neural network blocks but drift from from visual evidence in the image plane (see Fig. 3). That is to say, WHAM and GVHMR struggle to simultaneously deliver accurate per-frame pose and smooth, physically plausible motion with a unified output. In contrast, our method leverage the motion prior to enforce temporal consistency and the video-conditioned decoder to enforce the predicted motion to align with the visual cues, and both integrated to an end-to-end framework, producing one unified trajectory in the global space, achieving a good balance between motion realism and image alignment.

Qualitative results in Fig. 3 (row 1, 2) further demonstrate that our method yields substantially improved robustness under occlusions compared with the baselines.

Performance on occlusion-free videos. We further quantitatively evaluate MoRo in an occlusion-free scenario on RICH (Tab. 2). Our method delivers comparable results as baselines while yielding more plausible motion (lower Accel, G-Accel, and Jitter). Fig. 3 (row 2) illustrates that our reconstructions align closely with the video input.

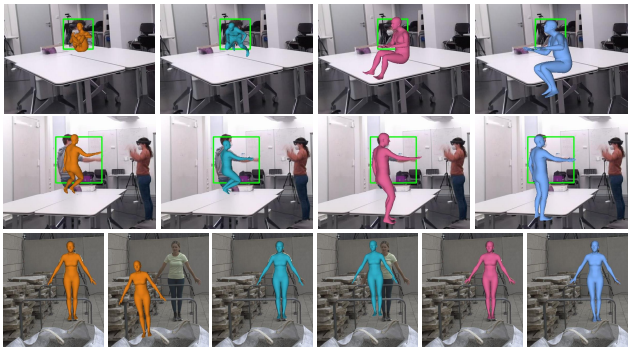

Figure 3. **Qualitative examples on EgoBody (row 1, 2) and RICH (row 3)**. For row 1-2, from left to right corresponds to WHAM-Cam, GVHMR-Cam, PromptHMR and ours. For row 3, from left to right corresponds to WHAM-Cam, WHAM-World, GVHMR-Cam, GVHMR-World, PromptHMR and ours.

5.6. Ablation Studies

We conduct ablation studies on EgoBody-occ to examine the impact of key design choices and the number of inference steps. As shown in Tab. 3, both motion realism and pose accuracy for occluded body parts improve notably when incorporating the motion prior ('ours' vs. 'w/o ME'), highlighting its effectiveness. The motion encoder (Sec. 4.2) not only enables masked modeling but, when pretrained on MoCap data, also better captures natural motion dynamics. The confidence-guided masking strategy (Sec. 4.5) further narrows the train–test gap, improving model's robustness

[†]While PromptHMR proposes a video-based variant (PromptHMR-Vid), the code for reproducing its result on RICH is unavailable. Thus we only evaluate the per-frame model. We will provide evaluation for PromptHMR-Vid in future versions once the evaluation code is provided.

	Method	PA-MPJPE↓	MPJPE↓ *-all*	MPJPE↓ *-vis*	MPJPE↓ *-occ*	PVE↓	GMPJPE↓	RTE↓	Accel↓	G-accel↓	Jitter↓	Sliding↓
per-frame	MEGA [17]	37.80	86.92	83.98	108.17	106.13	-	-	-	-	-	-
	TokenHMR [14]	38.07	76.38	72.94	101.30	94.71	-	-	-	-	-	-
	PromptHMR [67]	<u>35.00</u>	<u>48.34</u>	<u>45.30</u>	<u>70.42</u>	<u>59.79</u>	-	-	-	-	-	-
temporal-based	RoHM [79]	54.53	79.01	75.85	101.7	105.18	<u>308.8</u>	<u>2.23</u>	<u>2.81</u>	3.78	12.74	3.28
	WHAM [59] -Cam	44.20	82.03	77.33	116.1	98.46	745.23	5.18	3.27	115.68	626.52	72.78
	WHAM [59] -World	44.21	95.26	91.20	124.64	116.17	739.49	3.98	3.19	<u>3.36</u>	<u>10.27</u>	**2.73**
	GVHMR [57] -Cam	48.85	71.00	64.95	114.87	83.60	877.26	3.53	3.53	81.76	441.84	52.49
	GVHMR [57] -World	48.85	73.33	67.40	116.29	86.13	875.45	3.06	4.11	5.62	25.11	3.81
	Ours	**26.72**	**39.13**	**37.83**	**48.53**	**50.25**	**129.22**	**0.58**	**2.21**	**2.15**	**4.60**	<u>3.21</u>

Table 1. **Quantitative evaluation results on EgoBody-occ.** The best / second best results are in **boldface**, and <u>underlined</u>, respectively. For per-frame methods, we report only pelvis-aligned accuracy metrics since they lack global or temporal modeling by design.

	Method	PA-MPJPE↓	MPJPE↓	PVE↓	GMPJPE↓	RTE↓	Accel↓	G-Accel↓	Jitter↓	Sliding↓
per-frame	MEGA [17]	50.53	108.27	122.43	-	-	-	-	-	-
	TokenHMR [14]	40.37	77.74	90.68	-	-	-	-	-	-
	PromptHMR [67]	<u>38.17</u>	**58.56**	**67.25**	-	-	-	-	-	-
temporal-based	WHAM [59] -Cam	44.53	79.80	90.65	557.06	2.94	5.67	49.41	258.94	33.29
	WHAM [59] -World	44.53	102.58	117.07	660.60	4.40	5.49	6.51	21.01	<u>3.97</u>
	GVHMR [57] -Cam	39.78	<u>66.07</u>	<u>75.71</u>	<u>520.51</u>	2.42	<u>4.15</u>	17.36	83.92	14.57
	GVHMR [57] -World	39.78	73.72	83.85	553.66	<u>2.40</u>	4.40	<u>4.46</u>	<u>12.82</u>	**2.99**
	Ours	**35.37**	74.00	84.47	**378.43**	**1.45**	**3.71**	**3.59**	**5.90**	4.86

Table 2. **Quantitative evaluation results on RICH.** The best / second best results are in **boldface**, and <u>underlined</u>, respectively.

Method	MPJPE↓ *-vis*	MPJPE↓ *-occ*	GMPJPE↓	G-accel↓	Jitter↓	Sliding↓
Ours	**37.83**	**48.53**	**127.17**	**2.15**	**4.60**	**3.21**
w/o ME	39.77	55.89	134.31	3.66	14.20	5.92
w/o CGM	40.32	53.90	136.38	<u>2.24</u>	<u>5.13</u>	**3.21**
w/o $\mathcal{F}_{\text{smoother}}$	<u>38.66</u>	<u>49.50</u>	133.31	5.12	24.31	5.95
Inference steps						
T=1	38.27	51.10	129.22	2.32	6.12	3.57
T=5	**37.83**	<u>48.53</u>	**127.17**	**2.15**	4.60	3.21
T=10	<u>37.85</u>	**48.51**	<u>129.21</u>	**2.15**	**4.55**	<u>3.14</u>
T=20	37.92	48.65	129.52	**2.15**	<u>4.58</u>	**3.10**

Table 3. **Ablation study on EgoBody-occ** for the motion encoder ('ME', Sec. 4.2), pose token smoother ($\mathcal{F}_{\text{smoother}}$, Sec. 4.3), confidence-guided masking during training ('CGM', Sec. 4.5), and inference step numbers. The best / second best results are in **boldface**, and <u>underlined</u>, respectively.

during multi-step inference ('ours' vs. w/o 'CGM'). Finally, the pose token smoother (Sec. 4.3) mitigates motion jitters resulted from discrete quantization in the tokenizer ('ours' vs. w/o $\mathcal{F}_{\text{smoother}}$'). Study on inference steps reveals that, increasing inference steps consistently reduces jitter and foot sliding, improving overall motion realism. Meanwhile , global pose accuracy peaks at $T = 5$ steps, which we adopt as our final setup.

6. Conclusion

We introduced MoRo, a masked generative transformer framework for occlusion-robust human motion reconstruction from monocular video. MoRo leverages masked modeling and effectively consolidates multi-modal information across a set of heterogeneous datasets (MoCap, image-pose and video-motion data). By integrating a trajectory-aware motion prior and an image-conditioned pose prior into a video-conditioned generative transformer, MoRo recover temporally consistent human motion in global space in an end-to-end manner. Experiments show that MoRo outperforms state-of-the-art methods under occlusions while maintaining real-time performance, offering a practical solution for various downstream applications.

Limitations and future work. Despite its effectiveness, our method is currently restricted to static camera setups with known intrinsics, which limits its applicability to videos captured by moving cameras. In future work, we plan to incorporate techniques for modeling camera motion to extend MoRo to dynamic camera scenarios.

Acknowledgements. This work was supported as part of the Swiss AI initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project IDs #36 on Alps, enabling large-scale training. We sincerely thank Muhammed Kocabas for his help with the PromptHMR codebase, and Korrawe Karunratanakul for insightful discussions.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2, 4, 6

[2] 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. 2, 3, 6

[3] 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 *European Conference on Computer Vision (ECCV)*, 2016. 2

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. 2023. 3

[7] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In *ICCV*, 2019. 3

[8] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *ECCV*, 2022. 2

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 2

[10] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[11] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2, 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[14] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, 2024. 2, 3, 6, 8

[15] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3D human pose by watching humans in the mirror. In *CVPR*, 2021. 2

[16] Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Antonio Agudo, and Francesc Moreno-Noguer. Vq-hps: Human pose and shape estimation in a vector-quantized latent space. In *ECCV*, 2024. 2, 3

[17] Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Mega: Masked generative autoencoder for human mesh recovery, 2025. 2, 3, 6, 8

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

[19] 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. 2, 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 3

[21] 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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 6

[22] 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*, 2014. 2, 4, 6

[23] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[24] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[25] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2, 3

[26] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, 2022. 2

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[28] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2, 3

[29] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, 2021. 2, 3

[30] 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, 3, 6

[31] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[32] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[33] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 2, 3

[34] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 2

[35] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *CVPR*, 2021. 2

[36] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *ICCV*, 2023. 2

[37] Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Cewu Lu, Jan Kautz, and Umar Iqbal. Coin: Control-inpainting diffusion prior for human and camera motion estimation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[38] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2, 3, 4

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[40] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 4, 6

[41] Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3d human pose estimation. In *ECCV*, 2022. 2

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 2

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

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 3, 4, 6

[45] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. 2017. 2, 4, 6

[46] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 2

[47] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In *ICCV*, 2023. 2, 3

[48] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D vision (3DV)*, 2018. 2

[49] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. *arXiv preprint arXiv:2411.08128*, 2024. 3

[50] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[51] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Controlmm: Controllable masked motion generation. *arXiv preprint arXiv:2410.10780*, 2024. 3

[52] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, 2024. 2, 3

[53] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *ICCV*, 2021. 2, 3, 5

[54] Chris Rockwell and David Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 2

[55] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Pu Wang, Hongfei Xue, Srijan Das, and Chen Chen. Genhmr: Generative human mesh recovery, 2024. 2, 3, 6

[56] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems*, 37:140032–140065, 2024. 2

[57] 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 *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 3, 6, 7, 8

[58] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. PhaseMP: Robust 3D pose estimation via phase-conditioned human motion prior. In *ICCV*, 2023. 2, 3, 5

[59] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6, 7, 8

[60] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 2

[61] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5, 7

[62] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 6

[63] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. 3

[64] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[66] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 2, 6

[67] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 3, 6, 8

[68] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *CVPR*, 2022. 3

[69] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 2

[70] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 4

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

[72] Yingxuan You, Hong Liu, Ti Wang, Wenhao Li, Runwei Ding, and Xia Li. Co-evolution of pose and mesh for 3d human body estimation from video. In *ICCV*, 2023. 3

[73] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 2, 6

[74] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, 2021. 2

[75] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[76] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *ICCV*, 2021. 2, 3

[77] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 2, 6

[78] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *ICCV*, 2023. 2, 4

[79] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024. 1, 2, 3, 5, 6, 7, 8

[80] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan Yuille. 3d-aware neural body fitting for occlusion robust 3d human pose estimation. In *ICCV*, 2023. 2

[81] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[82] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, 2021. 2

[83] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 5