

# GenHMR: Generative Human Mesh Recovery

Muhammad Usama Saleem<sup>1</sup>, Ekkasit Pinyoanuntapong<sup>1</sup>, Pu Wang<sup>1</sup>, Hongfei Xue<sup>1</sup>, Srijan Das<sup>1</sup>,  
Chen Chen<sup>2</sup>

<sup>1</sup>University of North Carolina at Charlotte, Charlotte, NC, USA

<sup>2</sup>University of Central Florida, Orlando, FL, USA

{msaleem2, epinyoan, pwang13, hxue2, sdas24}@charlotte.edu, chen.chen@crcv.ucf.edu

## Abstract

Human mesh recovery (HMR) is crucial in many computer vision applications; from health to arts and entertainment. HMR from monocular images has predominantly been addressed by deterministic methods that output a single prediction for a given 2D image. However, HMR from a single image is an ill-posed problem due to depth ambiguity and occlusions. Probabilistic methods have attempted to address this by generating and fusing multiple plausible 3D reconstructions, but their performance has often lagged behind deterministic approaches. In this paper, we introduce GenHMR, a novel generative framework that reformulates monocular HMR as an image-conditioned generative task, explicitly modeling and mitigating uncertainties in the 2D-to-3D mapping process. GenHMR comprises two key components: (1) a pose tokenizer to convert 3D human poses into a sequence of discrete tokens in a latent space, and (2) an image-conditional masked transformer to learn the probabilistic distributions of the pose tokens, conditioned on the input image prompt along with randomly masked token sequence. During inference, the model samples from the learned conditional distribution to iteratively decode high-confidence pose tokens, thereby reducing 3D reconstruction uncertainties. To further refine the reconstruction, a 2D pose-guided refinement technique is proposed to directly fine-tune the decoded pose tokens in the latent space, which forces the projected 3D body mesh to align with the 2D pose clues. Experiments on benchmark datasets demonstrate that GenHMR significantly outperforms state-of-the-art methods.

**Project Website** — <https://m-usamasaleem.github.io/publication/GenHMR/GenHMR.html>

## Introduction

Recovering 3D human mesh from monocular images is an essential task in computer vision, with applications spanning diverse fields, such as character animation for video games and movies, metaverse, human-computer interaction, and sports performance optimization. However, recovering 3D human mesh from monocular images remains challenging due to inherent ambiguities in lifting 2D observations to 3D space, flexible body kinematic structures, complex intersections with the environment, and insufficient annotated 3D

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

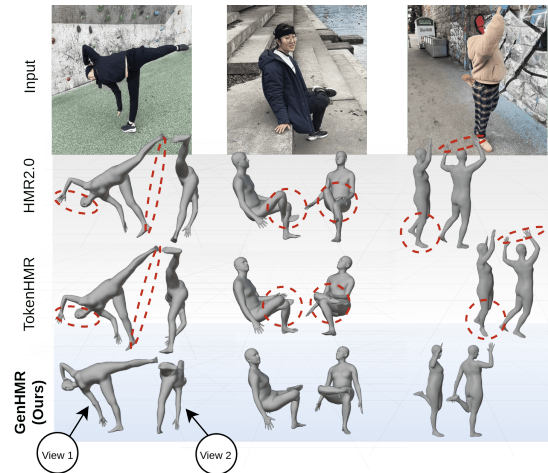


Figure 1: State of the art (SOTA) methods, HMR2.0 (Goel et al. 2023) and TokenHMR (Dwivedi et al. 2024), leverage vision transformers to recover 3D human mesh from a single image. The errors, highlighted by red circles, reveal the limitations of these SOTA approaches in handling unusual poses or ambiguous scenarios. Our method, GenHMR, overcomes these challenges by explicitly modeling and mitigating uncertainties in the 2D-to-3D mapping process, resulting in more accurate and robust 3D pose reconstructions in such complex scenarios.

data (Tian et al. 2023). To address these challenges, recent efforts have been focusing on two methods: (1) deterministic HMR and (2) probabilistic HMR. Both methods face critical limitations.

Deterministic methods, as the dominant approach for HMR, are designed to produce a single prediction for a given 2D image. These methods estimate the shape and pose parameters of 3D body model either by regressing from 2D image features extracted from deep neural networks (Choi et al. 2022; Kocabas et al. 2021; Kanazawa et al. 2018) or by directly optimizing the parametric body model by fitting it to 2D image cues, such as 2D keypoints (Bogo et al. 2016; Pavlakos et al. 2019; Xu et al. 2020), silhouettes (Omran et al. 2018), and part segmentations (Lassner et al. 2017). Recent deterministic HMR models, utilizing vision

transformers as their backbone, have achieved state-of-the-art (SOTA) mesh reconstruction accuracy. However, despite such promising progress, these SOTA deterministic models face a critical limitation: they constrain neural networks to produce a single prediction hypothesis. This approach overlooks the inherent depth ambiguity in monocular images, which can result in multiple plausible 3D reconstructions that equally fit the same 2D evidence.

To mitigate this problem, several works have proposed probabilistic approaches that generate multiple predictions from a single image using various generative models, such as conditional variational autoencoders (CVAEs) (Sharma et al. 2019; Sohn, Lee, and Yan 2015) and diffusion models (Shan et al. 2023; Holmquist and Wandt 2023). However, this increase in diversity typically comes at the cost of accuracy because strategical aggregation of multiple solutions into a single accurate prediction is challenging due to the potential kinematic inconsistency among these 3D human mesh predictions. As a result, none of these multi-hypothesis probabilistic methods are competitive with the latest single-output deterministic models.

To overcome the limitations of existing methods, we introduce GenHMR, a novel generative framework for 3D human mesh recovery from a single 2D image. GenHMR is built on two key components: a pose tokenizer and an image-conditional masked transformer. The framework follows a two-stage training paradigm. In the first stage, the pose tokenizer is trained using Vector Quantized Variational Autoencoders (VQ-VAE) (Van Den Oord, Vinyals et al. 2017), which convert the continuous human pose (i.e., the rotations of skeletal joints) into a sequence of discrete tokens in a latent space, based on a learned codebook. In the second stage, a portion of the pose token sequence is randomly masked. The image-conditional masked transformer is then trained to predict the masked tokens by learning the conditional categorical distribution of each token, given the input image and the unmasked tokens. This generative masking training allows GenHMR to learn an explicit probabilistic mapping from the 2D image to the human pose. Leveraging such feature, we propose uncertainty-guided iterative sampling during inference, where the model decodes multiple pose tokens simultaneously in each iteration by sampling from the learned image-conditioned pose distributions. The tokens with low prediction uncertainties are kept and the others are re-masked and re-predicted in the next iteration. This feature allows GenHMR to iteratively reduce 2D-to-3D mapping uncertainties and progressively correct the wrong joint rotations to improve the mesh reconstruction accuracy. To further refine the reconstruction quality, we propose a novel 2D pose-guided refinement technique, which directly optimizes the decoded pose tokens in the latent space, with an objective to force the projected 3D body mesh to align with the 2D pose clues. Our contributions are summarized as follows:

- We introduce GenHMR, a novel generative framework for accurate HMR from a single image, which largely departs from existing deterministic and probabilistic methods in terms of model architecture, training paradigm and

inference process.

- We leverage generative masking training to learn intricate image-conditioned pose distributions, thus effectively capturing the 2D-to-3D mapping uncertainty.
- We propose a novel iterative inference strategy, incorporating uncertainty-guided sampling followed by 2D pose-guided refinement to progressively mitigate HMR errors.
- We demonstrate through extensive experiments that GenHMR outperforms SOTA methods on standard datasets, including Human3.6M in the controlled environment, and 3DPW and EMDB for in-the-wild scenarios. For both cases, GenHMR could lead to 20 - 30 % error reduction (in terms of MPJPE) compared with SOTA methods. Qualitative results show that GenHMR is robust to ambiguous image observations (Fig. 1).

## Related Work

### Deterministic HMR

The field of Human Mesh Recovery (HMR) from monocular images has been primarily dominated by deterministic approaches, which aim to generate a single output for a given 2D image. The early work mainly adopts optimization-based approaches to fit a parametric model human model such as SMPL (Loper et al. 2015) to 2D image cues (Pavlakos et al. 2019; Kolotouros et al. 2019; Lassner et al. 2017). Later on, learning-based methods become more prevalent, which leverage CNNs to directly regress SMPL parameters from images (Choi et al. 2022; Kocabas et al. 2021; Kanazawa et al. 2018) and videos (Cho et al. 2023; Kanazawa et al. 2019). Recently, vision transformers (Alexey 2020) have been adopted for HMR tasks. For example, HMR 2.0 (Goel et al. 2023) and TokenHMR (Dwivedi et al. 2024) achieve the state-of-the-art mesh reconstruction accuracy by leveraging transformer’s ability of modeling the long-range correlations to learn the dependencies of different human body parts in HMR tasks. However, these deterministic methods are limited by their single-output nature, which fails to capture the inherent depth ambiguity in complex scenarios, leading to reconstruction errors when multiple plausible 3D interpretations exist.

### Probabilistic HMR

To address the limitations of deterministic methods, various probabilistic models have been exploited to address the inherent uncertainty in the reconstruction process and enable the generation of multiple plausible 3D mesh predictions from a single 2D image. These methods include mixture density networks (MDNs) (Bishop 1994; Li and Lee 2019), conditional variational autoencoders (CVAEs) (Sohn, Lee, and Yan 2015; Pavlakos et al. 2018), and normalizing flows (Wehrbein et al. 2021; Kolotouros et al. 2021). Recent advancements in diffusion-based HMR models, such as DDHPose (Cai et al. 2024), Diff-HMR (Cho and Kim 2023), and D3DP (Shan et al. 2023), have shown significant promise in generating diverse and realistic human meshes. However, these approaches face a great challenge in synthesizing multiple predictions into a single, coherent 3D pose.

Current methods often rely on simplistic aggregation techniques, such as averaging all hypotheses (Wehrbein et al. 2021; Li and Lee 2019, 2020) or performing pose or joint-level fusion (Shan et al. 2023; Cai et al. 2024), which frequently result in kinematically inconsistent and suboptimal reconstructions.

Unlike existing probabilistic methods that model and sample the distributions of entire 3D body meshes, GenHMR leverages masked generative transformers to model the pose and rotation distribution of each individual skeletal joint, while also capturing the inherent interdependence among joints. This approach enables our model to iteratively perform fine-grained joint-level sampling, progressively reconstructing the human mesh from ambiguous 2D image observations. Our generative framework draws inspiration from the great success of masked language and image models in text, image, and video generation (Ghazvininejad et al. 2019; Devlin et al. 2019; Zhang et al. 2021; Chang et al. 2022; Ding et al. 2022; Chang et al. 2023). However, while generative language and image models focus on increasing sample diversity, GenHMR aims to minimize 3D mesh generation diversity and uncertainty, conditioned on 2D image prompts.

## Proposed Method: GenHMR

The goal of GenHMR is to achieve accurate 3D human mesh reconstructions from monocular images  $I$  by learning body pose  $\theta$ , shape  $\beta$ , and camera parameters  $T$ . As shown in Figure 2, GenHMR comprises two main modules: the pose tokenizer and the image-conditioned masked transformer. The pose tokenizer converts 3D human pose parameters  $\theta$  into a sequence of discrete pose tokens. The image-conditioned masked transformer predicts masked pose tokens based on multi-scale image features extracted by an image encoder. During inference, we use an iterative decoding process to predict high-confidence pose tokens, progressively refining predictions by masking low-confidence tokens and leveraging both image semantics and inter-token dependencies. To further enhance accuracy, a 2D pose-guided sampling strategy is proposed to optimize the pose queries by aligning the re-projected 3D pose with the estimated 2D pose. Additionally, the shape parameters  $\beta$  and weak perspective camera parameters  $T$  are directly regressed from the image features, completing the 3D human mesh reconstruction process.

### Body Model

We utilize SMPL, a differentiable parametric body model (Loper et al. 2015), to represent the human body. The SMPL model encodes the body using pose parameters  $\theta \in \mathbb{R}^{24 \times 3}$  and shape parameters  $\beta \in \mathbb{R}^{10}$ . The pose parameters  $\theta = [\theta_1, \dots, \theta_{24}]$  include the global orientation  $\theta_1 \in \mathbb{R}^3$  of the whole body and the local rotations  $[\theta_2, \dots, \theta_{24}] \in \mathbb{R}^{23 \times 3}$  of the body joints, where each  $\theta_k$  represents the axis-angle rotation of joint  $k$  relative to its parent in the kinematic tree. Given these pose and shape parameters, the SMPL model generates a body mesh  $M(\theta, \beta) \in \mathbb{R}^{3 \times N}$ , where  $N = 6890$  vertices. The body joints  $J \in \mathbb{R}^{3 \times k}$  are then defined as a linear combination of these vertices, calculated using  $J =$

$MW$ , where  $W \in \mathbb{R}^{N \times k}$  represents fixed weights that map vertices to joints.

### Pose Tokenizer

The goal of the pose tokenizer is to learn a discrete latent space for 3D pose parameters by quantizing the encoder’s output into learned codebook  $C$ , as shown in Figure 2(a). We leverage the VQ-VAE (Van Den Oord, Vinyals et al. 2017) to pretrain the pose tokenizer. Specifically, given the SMPL pose parameters  $\theta$ , we use a convolution encoder  $E$  to map the pose parameters  $\theta$  into a latent embedding  $z$ . Each embedding  $z$  is then quantized to codes  $c \in C$  by finding the nearest codebook entry based on the Euclidean distance, described by  $\hat{z}_i = \arg \min_{c_k \in C} \|z_i - c_k\|_2$ . Then, the total loss function is defined as follows

$$\mathcal{L}_{\text{vq}} = \lambda_{\text{re}} \mathcal{L}_{\text{re}} + \lambda_{\text{E}} \| \text{sg}[z] - c \|_2 + \lambda_{\alpha} \| z - \text{sg}[c] \|_2$$

which consists of a SMPL reconstruction loss, a latent embedding loss and a commitment loss, where  $\lambda_{\text{re}}$ ,  $\lambda_{\text{E}}$ , and  $\lambda_{\alpha}$  are their respective weights.  $\text{sg}[\cdot]$  represents the stop-gradient operator. To improve reconstruction quality, we employ an L1 loss  $\mathcal{L}_{\text{re}} = \lambda_{\theta} \mathcal{L}_{\theta} + \lambda_{\text{V}} \mathcal{L}_{\text{V}} + \lambda_{\text{J}} \mathcal{L}_{\text{J}}$  to minimize the difference between the SMPL parameters and their ground-truth, including pose parameters  $\theta$ , mesh vertices  $V$ , and kinematic joints  $J$ . This tokenizer is optimized using a straight-through gradient estimator, with the codebooks being updated via exponential moving average and codebook reset, following the methodology outlined in (Esser, Rombach, and Ommer 2021; Williams et al. 2020).

### Image Conditioned Masked Transformer

The image conditioned masked transformer comprises two main components: the image encoder and the masked transformer decoder with multi-scale deformable cross attention.

**Image Encoder.** Our encoder employs a vision transformer (ViT) to extract image features (Alexey 2020; Dwivedi et al. 2024; Goel et al. 2023). We utilize the ViT-H/16 variant, which processes 16x16 pixel patches through transformer layers to generate feature tokens. Inspired by ViT-Det (Alexey 2020), we adopt a multi-scale feature approach by upsampling initial feature map from encoder to create a set of feature maps with varying resolutions. High-resolution feature maps capture fine-grained visual details (e.g., the presence and rotation of individual joints), while low-resolution feature maps preserve high-level semantics, e.g., the structure of the human skeleton.

**Masked Transformer Decoder.** Our decoder employs a multi-layer transformer whose inputs are the pose token sequence obtained from the pose tokenizer. These pose tokens serve as the queries that are cross-attended to the multi-scale feature maps from the image encoder. Since these feature maps are of high resolution, we adopt multi-scale deformable cross-attention to mitigate computational cost (Zhu et al. 2020). In particular, each pose token is only attended to a small set of sampling points around a reference point on multi-scale feature maps, regardless of the spatial size of the feature maps. The multi-scale deformable attention is expressed as:

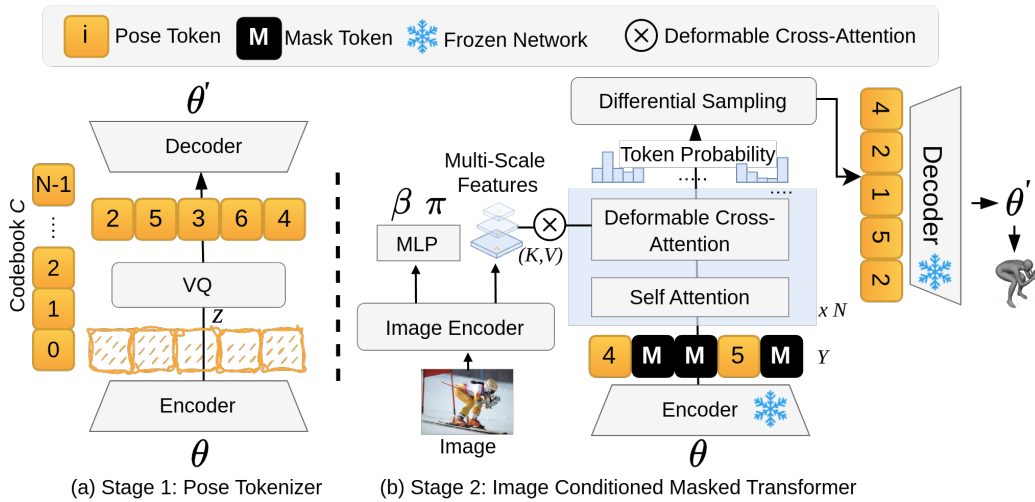


Figure 2: GenHMR *Training Phase*. GenHMR consists of two key components: (1) a **Pose Tokenizer** that encodes 3D human poses into a sequence of discrete tokens within a latent space, and (2) an **Image-Conditioned Masked Transformer** that models the probabilistic distributions of these tokens, conditioned on the input image and a partially masked token sequence.

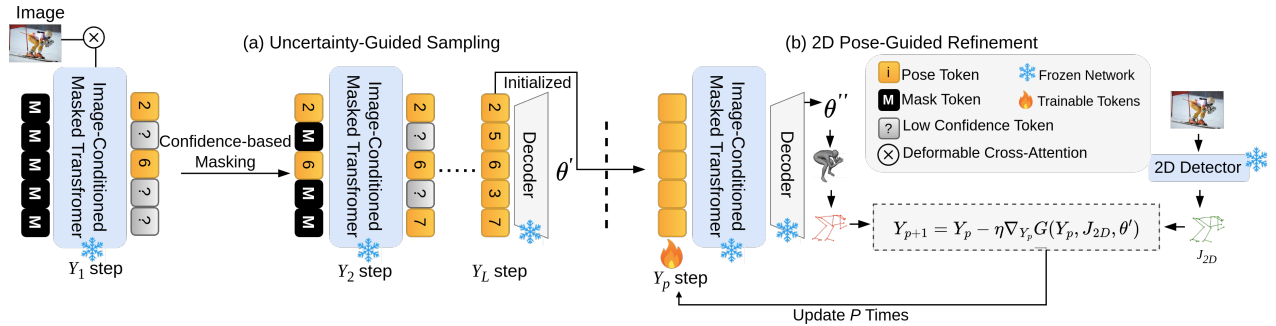


Figure 3: Our *inference strategy* comprises two key stages: (1) **Uncertainty-Guided Sampling**, which iteratively samples high-confidence pose tokens based on their probabilistic distributions, and (2) **2D Pose-Guided Refinement**, which fine-tunes the sampled pose tokens to further minimize 3D reconstruction uncertainty by ensuring consistency between the 3D body mesh and 2D pose estimates.

$$\text{MSDA}(Y, \hat{p}_y, \{x^l\}_{l=1}^L) = \sum_{l=1}^L \sum_{k=1}^K A_{l y k} \cdot \mathbf{W} x^l (\hat{p}_y + \Delta p_{l y k})$$

where  $Y$  represents pose token queries,  $\hat{p}_y$  denotes learnable reference points,  $\Delta p_{l y k}$  are the learnable sampling offsets around these points,  $\{x^l\}_{l=1}^L$  are the multi-scale image features,  $A_{l y k}$  are the attention weights, and  $\mathbf{W}$  is a learnable weight matrix. The [MASK] token is a special-purpose token used in the masked transformer decoder. During training, it represents masked pose tokens, and the model learns to predict the actual tokens to replace them. During inference, [MASK] tokens serve as placeholders for pose token generation.

### Training Strategy

**Generative Masking.** Given a pose token sequence  $Y = [y_i]_{i=1}^L$  from the pose tokenizer where  $L$  denotes the sequence length, our model is trained to reconstruct the pose

token sequence, conditioned on the image prompt under random masking strategies. In particular, we randomly mask out  $m = \lceil \gamma(\tau) \cdot L \rceil$  tokens, where  $\gamma(\tau) \in [0, 1]$  is a masking ratio function with  $\tau$  following a uniform distribution  $U(0, 1)$ . We adopt a cosine masking ratio function  $\gamma(\tau) = \cos(\frac{\pi\tau}{2})$  similar to the ones from generative text-to-image models (Chang et al. 2022, 2023). The masked tokens are replaced with learnable [MASK] tokens, forming the corrupted pose sequence  $Y_M$ . The categorical distribution of each pose token, conditioned on corrupted sequence  $Y_M$  and image prompt  $X$  is  $p(y_i|Y_M, X)$ , which explicitly models the uncertainty during the 2D-to-3D mapping process. The training objective is to minimize the negative log-likelihood of the pose token sequence prediction:

$$\mathcal{L}_{\text{mask}} = -\mathbb{E}_{Y \in \mathcal{D}} \left[ \sum_{\forall i \in [1, L]} \log p(y_i|Y_M, X) \right]. \quad (1)$$

**Training-time Differentiable Sampling.** The training loss,  $\mathcal{L}_{\text{mask}}$ , not only aids in capturing the uncertainty inherent in monocular human mesh recovery but also enforces accurate estimation of the pose parameter  $\theta$  within the discrete latent space. However, prior research (Kanazawa et al. 2018) indicates that, in addition to ensuring correct  $\theta$  estimation, it is highly advantageous to incorporate an additional 3D loss between the predicted and ground-truth 3D joints, as well as a 2D loss between the projections of these predicted 3D joints and the ground-truth 2D joints. The challenge in incorporating these losses into generative model training lies in the need to convert pose tokens in the latent space into the pose parameter  $\beta$  in the SMPL space. This conversion requires sampling the categorical distribution of pose tokens during training, which is non-differentiable. To overcome this challenge, we adopt the straight-through Gumbel-Softmax technique (Jang, Gu, and Poole 2016), which uses categorical sampling during the forward pass and employs differentiable sampling according to the continuous Gumbel-Softmax distribution during the backward pass, which can approximate the categorical distribution via temperature annealing. The final overall loss is  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{SMPL}} + \mathcal{L}_{3\text{D}} + \mathcal{L}_{2\text{D}}$ , which combines pose token prediction loss ( $\mathcal{L}_{\text{mask}}$ ), 3D loss ( $\mathcal{L}_{3\text{D}}$ ), 2D loss ( $\mathcal{L}_{2\text{D}}$ ), and SMPL parameter loss  $\mathcal{L}_{\text{SMPL}}$  that minimize the shape and pose parameters in the SMPL space.

### Inference Strategy

As shown in Fig. 3, our inference strategy comprises two key stages: (1) uncertainty-guided sampling, which iteratively samples high-confidence pose tokens based on their probabilistic distributions and (2) 2D pose-guided refinement, which fine-tunes the sampled pose tokens to further minimize 3D reconstruction uncertainty by ensuring the consistency between the 3D body mesh and 2D pose estimates.

**Uncertainty-Guided Sampling.** The sampling process begins with a fully masked sequence  $Y_1$  of length  $L$ , where all tokens are initially set to `[MASK]`. The sequence is decoded over  $T$  iterations. At each iteration,  $t$ , the masked tokens are decoded by performing stochastic sampling, where the tokens are sampled based on their prediction distributions  $p(y_i|Y_M, X)$ . After the token sampling, a certain number of tokens with low prediction confidences are re-masked and re-predicted in the next iteration. The number of tokens to be re-masked is determined by a masking schedule  $\lceil \gamma (\frac{t}{T}) \cdot L \rceil$ , where  $\gamma$  is a decaying function of iteration  $t$  that produces higher masking ratio in the early iterations when the prediction confidence is low, while yielding low masking ratio in the latter iterations when the prediction confidence increases as more context information becomes available from previous iterations. We adopt the cosine function for  $\gamma$  and the impact of other decaying functions is shown in the supplementary material.

**2D Pose-Guided Refinement.** To further reduce uncertainties and ambiguities in the 3D reconstruction, we refine the pose tokens  $Y$ , while keeping the whole network frozen, so that the 3D pose estimates are better aligned with 2D pose clues from off-the-shelf 2D pose detectors such as OpenPose (Cao et al. 2017). This optimization process is initialized by the pose tokens from uncertainty-guided sampling,

and these tokens are then iteratively updated to minimize a composite guidance function  $G(Y_p, J_{2\text{D}}, \theta')$  that penalizes the misalignment of 3D and 2D poses along with regularization terms:

$$Y^+ = \arg \min_{Y_p} (\mathcal{L}_{2\text{D}}(J'_{3\text{D}}) + \lambda_{\theta'} \mathcal{L}_{\theta'}(\theta')) \quad (2)$$

The term  $\mathcal{L}_{2\text{D}}(J'_{3\text{D}})$  ensures that the reprojected 3D joints  $J'_{3\text{D}}$  are aligned with the detected 2D keypoints  $J_{2\text{D}}$ :

$$\mathcal{L}_{2\text{D}}(J'_{3\text{D}}) = |\Pi(K(J'_{3\text{D}})) - J_{2\text{D}}|^2 \quad (3)$$

where  $\Pi(K(\cdot))$  represents the perspective projection with camera intrinsics  $K$ . The regularization term  $\mathcal{L}_{\theta'}(\theta')$  ensures that the pose parameters  $\theta'$  remain close to the initial estimate, preventing excessive deviations and maintaining plausible human body poses. At each iteration  $p$ , the pose embeddings  $Y_p$  are updated using the following gradient-based approach:

$$Y_{p+1} = Y_p - \eta \nabla_{Y_p} G(Y_p, J_{2\text{D}}, \theta') \quad (4)$$

Here,  $\eta$  controls the magnitude of the updates to the pose embeddings, while  $\nabla_{Y_p} G(Y_p, J_{2\text{D}}, \theta')$  represents the gradient of the objective function with respect to the pose embeddings  $Y$  at iteration  $p$ . This refinement process continues over  $P$  iterations and our experiments show that only a small number of iterations (5 to 10) is sufficient to yield satisfactory enhancement.

## Experiments

**Datasets.** We trained the pose tokenizer using the AMASS (Mahmood et al. 2019) standard training split and MOYO (Tripathi et al. 2023). For GenHMR, following prior work (Goel et al. 2023) and to ensure fair comparisons, we used standard datasets (SD): Human3.6M (H36M) (Ionescu et al. 2013), COCO (Lin et al. 2014), MPI-INF-3DHP (Mehta et al. 2017), and MPII (Andriluka et al. 2014).

**Evaluation Metrics.** We evaluate GenHMR using the Mean Per Joint Position Error (MPJPE) and Mean Vertex Error (MVE) for 3D pose estimation accuracy. We also report the Procrustes-Aligned MPJPE (PA-MPJPE) to assess the alignment between predicted and ground-truth poses after rigid transformation. To quantify computational efficiency, we use Average Inference Time per Image (AITI) (s), similar to the average inference/optimization time per image reported by OpenPose (Cao et al. 2017). However, AITI measures per-iteration processing time, enabling fine-grained analysis of our uncertainty-guided sampling and 2D Pose-guided refinement. Lower values across all these metrics indicate better performance. GenHMR is tested on the Human3.6M testing split, following previous works (Kolotouros et al. 2019). To evaluate GenHMR’s generalization on challenging in-the-wild datasets with varying camera motions and diverse 3D poses, we test 3DPW (Von Marcard et al. 2018) and EMDB (Kaufmann et al. 2023) without training on them, ensuring a fair assessment on unseen data.

### Comparison to State-of-the-art Approaches

We evaluate our approach against a range of state-of-the-art deterministic and probabilistic HMR methods on the Human3.6M, 3DPW, and EMDB datasets, as detailed in Table 1. GenHMR consistently outperforms existing methods

Methods	Venue	Human3.6M		3DPW		MVE	EMDB		MVE
		PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)	MPJPE (↓)		PA-MPJPE (↓)	MPJPE (↓)	
<b>Deterministic HMR Methods</b>									
FastMETRO (Cho 2022)	<i>ECCV 2022</i>	33.7	52.2	65.7	109.0	121.6	72.7	108.1	119.2
PARE (Kocabas et al. 2021)	<i>ICCV 2021</i>	50.6	76.8	50.9	82.0	97.9	72.2	113.9	133.2
Virtual Marker (Ma et al. 2023)	<i>CVPR 2023</i>	-	-	48.9	80.5	93.8	-	-	-
CLIFF (Li et al. 2022)	<i>ECCV 2022</i>	<u>32.7</u>	47.1	46.4	73.9	87.6	68.8	103.1	122.9
HMR2.0 (Goel et al. 2023)	<i>ICCV 2023</i>	33.6	<u>44.8</u>	<u>44.5</u>	<u>70.0</u>	<u>84.1</u>	<u>61.5</u>	<u>97.8</u>	120.1
VQ HPS (Fiche et al. 2023)	<i>ECCV 2024</i>	-	-	45.2	71.1	84.8	65.2	99.9	<u>112.9</u>
TokenHMR (Dwivedi et al. 2024)	<i>CVPR 2024</i>	36.3	48.4	47.5	75.8	86.5	66.1	98.1	116.2
<b>Probabilistic HMR Methods</b>									
Diff-HMR (Cho and Kim 2023) <sup>†</sup>	<i>ICCV 2023</i>	-	-	55.9	94.5	109.8	-	-	-
3D Multibodies (Biggs et al. 2020) <sup>†</sup>	<i>NeurIPS 2020</i>	42.2	58.2	55.6	75.8	-	-	-	-
ProHMR (Kolotouros et al. 2021) <sup>†</sup>	<i>ICCV 2021</i>	-	-	52.4	84.0	-	-	-	-
<b>GenHMR</b>	<b>Ours</b>	<b>22.4 (10.3↓)</b>	<b>33.5 (11.3↓)</b>	<b>32.6 (13.8↓)</b>	<b>54.7 (15.3↓)</b>	<b>67.5 (17.3↓)</b>	<b>38.2 (23.3↓)</b>	<b>68.5 (31.4↓)</b>	<b>76.4 (43.7↓)</b>

Table 1: Reconstructions Evaluated in 3D: Reconstruction errors (in mm) on the Human3.6M, 3DPW, and EMDB datasets. Lower values (↓) indicate better performance. Underlined results show the second-best performance in each column. **Blue** indicates improvements of our method compared to the second-best method. & – & means results are not reported. <sup>†</sup>Results for existing probabilistic HMR methods are reported for 25 multiple hypotheses.

across key evaluation metrics—PA-MPJPE, MPJPE, and MVE—demonstrating its superior ability to produce accurate 3D reconstructions. A significant contributor to this success is GenHMR’s capability to model and refine uncertainty throughout the reconstruction process, making it particularly effective in challenging scenarios involving complex poses and occlusions.

This capability enables GenHMR to not only achieve substantial performance gains on the controlled Human3.6M dataset but also to deliver impressive error reduction on the more challenging in-the-wild 3DPW and EMDB datasets. For example, GenHMR achieves a 25.2% reduction in MPJPE on the Human3.6M dataset, a 21.8% reduction on 3DPW, and a remarkable 29.9% reduction on EMDB, compared to existing state-of-the-art methods. These consistent improvements across multiple datasets, even without training on in-the-wild datasets like 3DPW and EMDB, highlight GenHMR’s effectiveness in addressing the complexities of real-world scenarios and delivering high-quality 3D reconstructions with unprecedented accuracy.

## Ablation Study

The key to GenHMR’s effectiveness lies in its mask modeling and iterative refinement techniques. In this ablation study, we investigate how iterative refinement and mask-scheduling strategies influence the model’s performance. In the **Supplementary Material**, we provide extensive additional experimental results and visualizations that offer in-depth analyses of key factors contributing to GenHMR’s performance. These include: (1) architectural components (pose tokenizer design, backbones, feature resolutions), (2) training strategies (masking scheduling, Gumbel-Softmax annealing, regularization via keypoint and SMPL losses in GenHMR), (3) inference techniques (speed-accuracy trade-offs in iterative refinement stages), (4) impact of training dataset size, and (5) model limitations.

**Effectiveness of Uncertainty-Guided Sampling.** The number of iterations during inference plays a critical role in balancing speed and accuracy in 3D pose estimation. As

# of iter.	AITI (sec)	3DPW		EMDB	
		MPJPE	MVE	MPJPE	MVE
1	0.032	73.5	84.2	92.2	104.5
3	0.075	70.2	81.9	89.8	101.6
5	0.102	68.1	77.5	88.2	99.5
10	0.255	67.8	79.1	87.5	98.3
15	0.321	67.6	78.5	87.1	96.6
20	0.412	67.4	78.1	86.8	95.4

Table 2: Impact of Iterations in Uncertainty-Guided Sampling

# of iter.	H36M	3DPW		EMDB		AITI (s)
	MPJPE	MPJPE	MVE	MPJPE	MVE	
UGS*	37.1	68.1	77.5	88.2	99.5	0.102
1	36.2	66.5	77.0	86.5	96.5	0.154
3	35.1	63.6	73.5	83.7	92.1	0.209
5	34.3	60.2	70.4	79.1	87.5	0.253
10	33.2	57.5	68.5	73.5	82.0	0.444
20	33.5	54.7	67.5	68.5	76.4	0.638

Table 3: Impact of iterations on 2D Pose Guided Refinement. The first row UGS\* reflects Uncertainty-Guided Sampling (UGS) with 5 iterations used to initialize the pose query  $Y$ . AITI is obtained on a single mid-grade GPU (NVIDIA RTX A5000).

demonstrated in Table 2, increasing the iterations generally enhances the accuracy of pose reconstructions, as reflected by improvements in MVE and MPJPE. For instance, on the 3DPW dataset, MVE decreases from 84.2 to 78.1, and MPJPE drops from 73.5 to 67.4 when the number of iterations increases from 1 to 20. However, beyond a certain threshold, such as after 10 iterations, the benefits of additional iterations diminish, with only marginal reductions in error. Notably, even with a smaller number of iterations, like 5, the model achieves significant accuracy improvements.

**Effectiveness of 2D Pose Guided Refinement.** Fig. 4 visualizes that 2D pose-guided refinement can iteratively improve 3D pose reconstruction accuracy particularly in sce-



Figure 4: Impact of 2D Pose-Guided Refinement on 3D pose reconstruction. Red circles highlight areas of errors after each refinement iteration, which showcases how the method progressively refines these poses. By fine-tuning pose tokens to align the 3D pose with 2D detections, our method iteratively reduces uncertainties and improves accuracy. Significant improvements are seen in the early iterations, with errors largely minimized at the 10th iteration. Note that the initial mesh comes from UGS.

narios with complex poses and occlusions. As shown in Table 3, significant accuracy gains occur early: on 3DPW, MVE drops by 11.04% (77.0 mm to 68.5 mm) from 1 to 10 iterations, with minimal improvement to 67.5 mm at 20 iterations, where MVE improves by 12.34% from 1 to 20 iterations. On EMDB dataset, MVE metric decreases by 15.02% (96.5 mm to 82.0 mm) from 1 to 10 iterations, further reducing to 20.8% (76.4 mm) at 20 iterations. These results underscore the method’s effectiveness, though a trade-off with computational time is evident, as AITI increases from 0.154 seconds at 1 iteration to 0.638 seconds at 20 iterations. Notably, substantial gains are achievable with fewer iterations; at 5 iterations, MVE improves by 8.57% on 3DPW (77.0 mm to 70.4 mm) and 9.3% on EMDB (96.5 mm to 87.5 mm), while AITI remains manageable at 0.253 seconds. This highlights that 2D Pose-Guided refinement effectively balances accuracy and efficiency, making it practical for real-time human mesh recovery.

Masking Ratio $\gamma(\tau)$	3DPW		EMDB	
	MPJPE	MVE	MPJPE	MVE
$\gamma(\tau \in \mathcal{U}(0, 1))$	68.1	77.5	<b>88.2</b>	99.5
$\gamma(\tau \in \mathcal{U}(0, 0.3))$	72.1	84.2	93.5	105.3
$\gamma(\tau \in \mathcal{U}(0, 0.5))$	69.8	81.7	90.2	101.5
$\gamma(\tau \in \mathcal{U}(0, 0.7))$	<b>67.9</b>	<b>77.1</b>	88.7	<b>99.4</b>

Table 4: Impact of masking ratio during training. These results were derived from the Uncertainty-Guided Sampling (UGS) stage with 5 iterations, where we evaluated the initial pose and shape estimate by iteratively refining the pose tokens.

**Masking Ratio during Training.** The ablation study on masking ratios during training underscores their influence on the accuracy of our generative model for human mesh recovery (HMR) as shown in Table 4. We utilize a cosine-

based masking ratio function  $\gamma(\tau)$ , with  $\tau$  sampled from a uniform distribution, to randomly mask segments of the pose token sequence during training. This method induces varying levels of information loss, compelling the model to develop more robust reconstruction capabilities. The results reveal that broader masking ratios, such as  $\gamma(\tau \in \mathcal{U}(0, 0.7))$ , lead to the most accurate reconstructions, demonstrated by the lowest MPJPE of 67.9 mm and a corresponding MVE of 77.1 mm on the 3DPW dataset. In contrast, narrower masking ratios, like  $\gamma(\tau \in \mathcal{U}(0, 0.3))$ , result in higher error rates, with MPJPE rising to 72.1 mm and MVE to 84.2 mm, suggesting reduced generalization capabilities. These findings emphasize the pivotal role of selecting an appropriate masking ratio during training to achieve effective model generalization and precise 3D pose reconstruction.

## Conclusion

In this paper, we introduced GenHMR, a novel generative approach to monocular human mesh recovery that effectively addresses the longstanding challenges of depth ambiguity and occlusion. By reformulating HMR as an image-conditioned generative task, GenHMR explicitly models and mitigates uncertainties in the complex 2D-to-3D mapping process. At its core, GenHMR consists of two key components: a pose tokenizer that encodes 3D human poses into discrete tokens within a latent space, and an image-conditional masked transformer that learns rich probabilistic distributions of these pose tokens. These learned distributions enable two powerful inference techniques: uncertainty-guided iterative sampling and 2D pose-guided refinement, which together produce robust and accurate 3D human mesh reconstructions. Extensive experiments on benchmark datasets demonstrate GenHMR’s superiority over SOTA HMR methods.

## References

- Alexey, D. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3686–3693.
- Biggs, B.; Novotny, D.; Ehrhardt, S.; Joo, H.; Graham, B.; and Vedaldi, A. 2020. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in neural information processing systems*, 33: 20496–20507.
- Bishop, C. M. 1994. Mixture density networks.
- Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, 561–578. Springer.
- Cai, Q.; Hu, X.; Hou, S.; Yao, L.; and Huang, Y. 2024. Disentangled Diffusion-Based 3D Human Pose Estimation with Hierarchical Spatial and Temporal Denoiser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 882–890.
- Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.
- Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.
- Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11315–11325.
- Cho, H.; Ahn, J.; Cho, Y.; and Kim, J. 2023. Video inference for human mesh recovery with vision transformer. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–6. IEEE.
- Cho, H.; and Kim, J. 2023. Generative approach for probabilistic human mesh recovery using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4183–4188.
- Cho, Y. K. O. T.-H., Junhyeong. 2022. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, 342–359. Springer.
- Choi, H.; Moon, G.; Park, J.; and Lee, K. M. 2022. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1475–1484.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902.
- Dwivedi, S. K.; Sun, Y.; Patel, P.; Feng, Y.; and Black, M. J. 2024. TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1323–1333.
- Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.
- Fiche, G.; Leglaive, S.; Alameda-Pineda, X.; Agudo, A.; and Moreno-Noguer, F. 2023. VQ-HPS: Human Pose and Shape Estimation in a Vector-Quantized Latent Space. *arXiv preprint arXiv:2312.08291*.
- Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.
- Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; and Malik, J. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14783–14794.
- Holmquist, K.; and Wandt, B. 2023. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15977–15987.
- Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.
- Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.
- Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.
- Kanazawa, A.; Zhang, J. Y.; Felsen, P.; and Malik, J. 2019. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5614–5623.
- Kaufmann, M.; Song, J.; Guo, C.; Shen, K.; Jiang, T.; Tang, C.; Zárate, J. J.; and Hilliges, O. 2023. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14632–14643.
- Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11127–11137.
- Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and

- shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2252–2261.
- Kolotouros, N.; Pavlakos, G.; Jayaraman, D.; and Daniilidis, K. 2021. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11605–11614.
- Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M. J.; and Gehler, P. V. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6050–6059.
- Li, C.; and Lee, G. H. 2019. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9887–9895.
- Li, C.; and Lee, G. H. 2020. Weakly supervised generative network for multiple 3d human pose hypotheses. *arXiv preprint arXiv:2008.05770*.
- Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; and Yan, Y. 2022. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, 590–606. Springer.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.
- Ma, X.; Su, J.; Wang, C.; Zhu, W.; and Wang, Y. 2023. 3d human mesh estimation from virtual markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 534–543.
- Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.
- Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.
- Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; and Schiele, B. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, 484–494. IEEE.
- Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985.
- Pavlakos, G.; Zhu, L.; Zhou, X.; and Daniilidis, K. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 459–468.
- Shan, W.; Liu, Z.; Zhang, X.; Wang, Z.; Han, K.; Wang, S.; Ma, S.; and Gao, W. 2023. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14761–14771.
- Sharma, S.; Varigonda, P. T.; Bindal, P.; Sharma, A.; and Jain, A. 2019. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2325–2334.
- Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.
- Tian, Y.; Zhang, H.; Liu, Y.; and Wang, L. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence*.
- Tripathi, S.; Müller, L.; Huang, C.-H. P.; Taheri, O.; Black, M. J.; and Tzionas, D. 2023. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4713–4725.
- Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.
- Wehrbein, T.; Rudolph, M.; Rosenhahn, B.; and Wandt, B. 2021. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11199–11208.
- Williams, W.; Ringer, S.; Ash, T.; MacLeod, D.; Dougherty, J.; and Hughes, J. 2020. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33: 4524–4535.
- Xu, H.; Bazavan, E. G.; Zangir, A.; Freeman, W. T.; Sukthankar, R.; and Sminchisescu, C. 2020. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6184–6193.
- Zhang, Z.; Ma, J.; Zhou, C.; Men, R.; Li, Z.; Ding, M.; Tang, J.; Zhou, J.; and Yang, H. 2021. M6-UFC: Unifying multi-modal controls for conditional image synthesis via non-autoregressive generative transformers. *arXiv preprint arXiv:2105.14211*.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *ArXiv*, abs/2010.04159.