RAY-FLOW HMR: PROBABILISTIC HUMAN MESH RECOVERY WITH ACCURATE, CONSISTENT, AND DIVERSE SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular human mesh recovery is an inherently ambiguous task. This motivates a probabilistic approach. Recent works learn the distribution of body meshes conditioned on the image. However, these methods often produce samples that lack accuracy, consistency, diversity, or some combination of these. We argue that learning image-consistent samples in body-model parameter space is difficult, since the mapping that goes through body geometry and perspective projection is complex and non-linear. In this work, we propose to enforce keypoint consistency at model construction, by parameterising 3D keypoints as depths along camera rays formed by 2D keypoints. Our model, Ray-Flow HMR, generates body samples in three stages: first, we sample 2D keypoints using an image-conditioned normalising flow; then, we sample their depths along the corresponding rays using another conditional normalising flow; finally, we regress body meshes via a learned inverse kinematics module. This construction ensures that sampled 3D keypoints project to the conditioning 2D keypoints, allowing for inverse kinematics to better preserve image-space consistency for the mesh samples. On in-the-wild datasets, Ray-Flow HMR outperforms probabilistic baselines in accuracy, consistency, and diversity, while remaining competitive with deterministic methods.

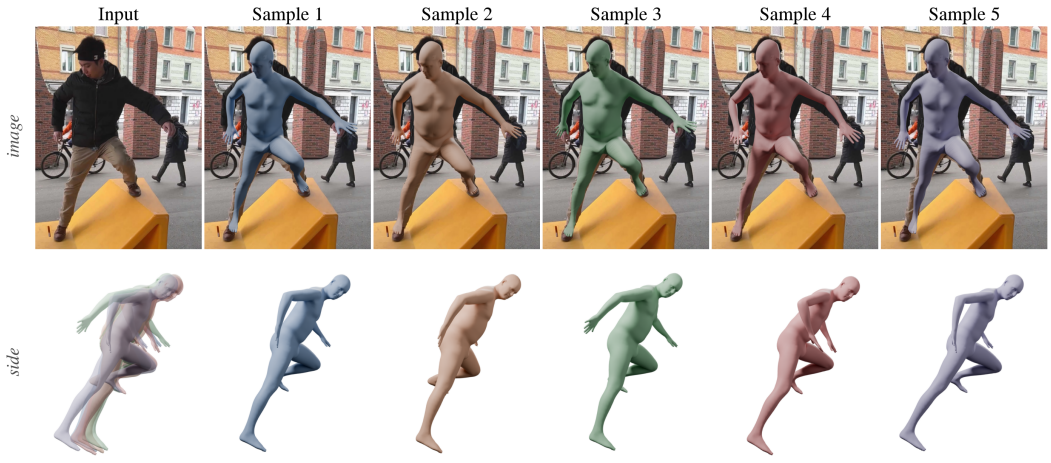

Figure 1: Samples from our Ray-Flow HMR model. We learn the conditional distribution of body meshes given an image. Our model generates high-quality samples that are consistent with the image, while being diverse in the ambiguous dimensions.

1 INTRODUCTION

Monocular human mesh recovery (HMR) is a long-standing computer vision task with a wide variety of applications in robotics, motion capture, sign-language parsing, movie and gaming industries,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

etc. It is also an inherently ambiguous problem: occluded or out-of-frame body parts, scale and depth ambiguity, body under loose clothing, camera resolution and scene lighting are all sources of ambiguity in an image. As a result, multiple 3D body meshes might explain a 2D image equally well. This motivates a probabilistic approach which seeks to represent the range of plausible bodies compatible with the available image evidence, instead of treating HMR as a deterministic problem.

Amongst recent advances, ProHMR (Kolotouros et al., 2021) models the distribution of SMPL parameters using a conditional normalising flow; HuManiFlow (Sengupta et al., 2023) factorises the pose distribution along the human joint kinematic chain; HuMR (Wehrbein et al., 2025) leverages the uncertainty in a 2D keypoint estimator to model the distribution in 3D. Importantly, HuManiFlow defines three metrics to assess the learned distribution: *sample accuracy*, *sample-input consistency*, and *sample diversity*. We observe that despite recent efforts, probabilistic methods exhibit a trade-off amongst the three metrics. This is further illustrated in Section 4, where we compare samples from these baselines.

The lack of consistency is particularly problematic: a plausible predictive distribution should generate samples that are well aligned with the image evidence, and without such consistency, we cannot convincingly argue that monocular ambiguity has been captured faithfully. We argue that the difficulty lies in accurately constraining the predictive distribution to the consistency manifold. This is defined as the set of body configurations that agree equally well with the image evidence when they are reprojected. More specifically, in this work, we discuss 2D keypoint reprojection, and the manifold is the set of body meshes that project to the same 2D keypoints. Commonly, HMR systems are supervised via a 2D reprojection loss (Kanazawa et al., 2018; Goel et al., 2023; Patel & Black, 2025), which attempts to constrain the model to this manifold, or a point on it in the deterministic case. However, there is no explicit guidance or constraint for the model. This is problematic since the mapping from model outputs to the manifold is complex and highly non-linear. Typically a model samples parameters from a body model such as SMPL (Loper et al., 2023). These samples go through the body model forward mapping, before being placed in camera space, and finally reprojected onto the image where the loss is calculated. Black-box models may struggle learn this distribution effectively. We expand on the complexity discussion in Section 3.

In this paper, we propose a factorised approach that solves the problem in three stages: an image-space normalising flow samples 2D skeletons, a conditional 3D keypoint normalising flow generates 3D hypotheses as depths along their corresponding camera rays, and finally a learned inverse kinematics decoder converts these samples into body meshes. This factorisation allows us to explicitly constrain the model predictive distribution to the consistency manifold. We term our model *Ray-Flow HMR* and evaluate it on in-the-wild image datasets 3DPW (Von Marcard et al., 2018) and EMDB (Kaufmann et al., 2023) against deterministic and probabilistic baselines. We show that through our explicit probabilistic factorisation, our model improves point and sample accuracy, sample-input consistency, and sample diversity.

In summary, we make the following contributions:

- We argue that the difficulty in faithfully modelling the image-conditioned predictive distribution lies in the lack of explicit representation of the image consistency manifold.
- We propose a factorised approach that progressively solves for 2D keypoints, 3D keypoints, and the mesh via inverse kinematics.
- We show that our model achieves higher accuracy, consistency, and diversity compared to existing probabilistic baselines.

2 Related Work

2.1 Deterministic human mesh recovery

Deterministic human mesh recovery aims to predict the most accurate mesh from an image without putting focus on understanding the ambiguity. Earlier methods use optimisation to fit a body mesh to image evidence (Bogo et al., 2016; Guan et al., 2009), using 2D keypoint detectors such as HRNet (Sun et al., 2019) and vision transformers (Dosovitskiy et al., 2020; Xu et al., 2022). HMR (Kanazawa et al., 2018) directly regresses a SMPL (Loper et al., 2023; Pavlakos et al., 2019) mesh using an end-to-end network, which set the baseline for a large corpus of follow-up works:

Kolotouros et al. (2019); Kocabas et al. (2021); Lin et al. (2021); Zhang et al. (2021); Li et al. (2021; 2022); Corona et al. (2022); Goel et al. (2023); Wang & Daniilidis (2023); Black et al. (2023); Dwivedi et al. (2024b); Cai et al. (2023); Wang et al. (2024); Patel & Black (2025); Yin et al. (2025); Sárándi & Pons-Moll (2024); Yang et al. (2026). Notably, HybrIK (Li et al., 2021) and NIKI (Li et al., 2023) uses inverse kinematics (IK) to decompose the problem; Ray3D (Zhan et al., 2022) explicitly models camera rays to constrain 3D pose predictions; CameraHMR (Patel & Black, 2025) leverages proper perspective projection to better supervise the reprojection error; Recent methods have scaled up their training datasets to demonstrate strong accuracy and generalisability (Cai et al., 2023; Yin et al., 2025; Yang et al., 2026). Deterministic methods do not provide any probabilistic interpretation of their predictions.

2.2 Probabilistic Human Mesh Recovery

Probabilistic HMR models the uncertainty within the image. Earlier methods used categorical distributions (Biggs et al., 2020), mixture density networks (Bishop, 1994; Li & Lee, 2019; Oikarinen et al., 2021), and others (Sengupta et al., 2021b; Fang et al., 2023; Dwivedi et al., 2024a; Saleem et al., 2025; Sengupta et al., 2021a; Fiche et al., 2025). More recently, normalising flow-based methods(Wehrbein et al., 2021; Kolotouros et al., 2021; Sengupta et al., 2023; Wehrbein et al., 2025) have demonstrated expressiveness in the learned distribution. ProHMR (Kolotouros et al., 2021) adapts the HMR network to output context for a normalising flow; HuManiFlow (Sengupta et al., 2023) explicitly models the human kinematic chain and formulates a hierarchical normalising flow stream; HuMR (Wehrbein et al., 2025) leverages the uncertainty in a 2D keypoint estimator to infer the 3D distribution. Despite these advances, probabilistic methods exhibit a trade-off amongst prediction accuracy, sample-input consistency, and sample diversity.

3 Methods

3.1 Preliminaries

SMPL. SMPL (Loper et al., 2023) is the de facto parametric body model used for human mesh recovery. Given a set of pose and shape parameters, SMPL defines a differentiable forward mapping that generates the human mesh:

$$\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \mathbf{V} \in \mathbb{R}^{6890\times 3}, \tag{1}$$

where $\boldsymbol{\theta} \in \mathrm{SO}(3)^{24}$ are pose parameters defined as a set of axis-angle rotation vectors, and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are shape parameters obtained via principal component analysis (PCA). SMPL also defines a linear regressor $\mathcal{J}$ that maps the mesh vertices to 3D joints $\mathbf{J} = \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\beta})$. We predict SMPL parameters in our method, as well as a camera-frame translation $\mathbf{t} \in \mathbb{R}^3$ that places both outputs in the camera coordinate system. We write the complete body state as $\mathbf{z} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$.

Normalising flows. Conditional normalising flows (Papamakarios et al., 2021) model expressive conditional distributions by transforming a simple base variable $Z \sim p_Z$ through an invertible mapping $X = f(Z; \mathbf{c})$ conditioned on context $\mathbf{c}$. Assuming the base distribution is independent of $\mathbf{c}$, the resulting density is

$$p_X(\mathbf{x} \mid \mathbf{c}) = p_Z\big(f^{-1}(\mathbf{x}; \mathbf{c})\big) \left| \det \frac{\partial f^{-1}(\mathbf{x}; \mathbf{c})}{\partial \mathbf{x}} \right|. \tag{2}$$

This formulation permits exact likelihood evaluation and direct sampling while allowing the context itself to be non-invertible. We use conditional flows to model distributions over 2D and 3D keypoints given image and camera features.

3.2 Ray-Flow HMR

Problem formulation. Given an image $\mathbf{I}$ of a person and its camera intrinsics $\mathbf{K}$, we model a conditional distribution $p(\mathbf{z} \mid \mathbf{I}, \mathbf{K})$ over body states. Through the SMPL forward mapping and translation, this induces a distribution over camera-frame body meshes. Our goal is to learn a distribution with high *consistency*: to generate samples that agree with the image, and high *diversity*: to generate diverse samples to represent plausible alternative reconstructions under the ambiguity of the input image.

The consistency manifold. We have argued in the introduction that constraining the predictive distribution to be consistent with the image is difficult. Here, we expand on this with a more formal analysis. Given the reprojection map $\Pi_{\mathbf{K},j}(\mathbf{z}) = \pi_{\mathbf{K}}\big(\mathcal{J}_j(\boldsymbol{\theta}, \boldsymbol{\beta}) + \mathbf{t}\big)$. For observed joint locations $\hat{\mathbf{u}}_{\mathcal{V}}$, where $\mathcal{V}$ denotes the set of visible joints, we define the consistency set

$$\mathcal{A}_{\mathbf{K},\mathcal{V}}(\hat{\mathbf{u}}) = \{\mathbf{z} : \Pi_{\mathbf{K},j}(\mathbf{z}) = \hat{\mathbf{u}}_j, \quad j \in \mathcal{V}\}. \tag{3}$$

Moving within this set changes the reconstructed body while preserving its visible joint projections. In practice, keypoint localisation uncertainty broadens the exact set into a neighbourhood around it. At a regular point $\mathbf{z}$, the tangent space of the set is the null space of the Jacobian of the reprojection with respect to the body states

$$T_{\mathbf{z}}\mathcal{A}_{\mathbf{K},\mathcal{V}} = \ker \mathbf{J}_{\Pi}(\mathbf{z}), \qquad \mathbf{J}_{\Pi}(\mathbf{z}) = \frac{\partial \Pi_{\mathbf{K},\mathcal{V}}(\mathbf{z})}{\partial \mathbf{z}}. \tag{4}$$

The Jacobian needs to differentiate through the perspective projection and the SMPL model, both of which are non-linear. This is further complicated by the body model, as SMPL joints are non-linearly determined by its pose and shape parameters. We argue that without explicit guidance, a black-box model will struggle to learn to traverse along these projection null directions effectively, resulting in inconsistency when samples fall off the null space, and lack of diversity if the space is insufficiently explored.

Leveraging camera rays to enforce consistency. Instead of solely relying on reprojection losses to implicitly supervise consistency, we propose to solve the easier 2D problem first, and constrain the sample space of the 3D problem to the 2D results. Figure 2 provides an intuition of our method. More specifically, we enforce model consistency by building rays from the camera optical centre to predicted 2D keypoint samples, and modelling 3D keypoints as depths along these rays. This guarantees 3D keypoint consistency with the 2D results.

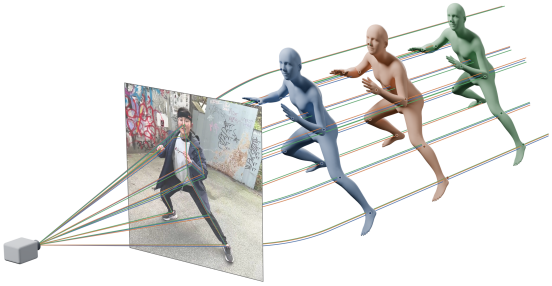

Figure 2: **Method overview.** Sampled 2D keypoints define camera rays, along which we sample depths to obtain diverse, image-consistent 3D hypotheses; a learned IK decoder converts these hypotheses into body meshes.

Mathematically, given a predicted 2D keypoint $\mathbf{u}_j \in \mathbb{R}^2$, we constrain its 3D location to the camera ray $\mathbf{X}_j = \ell_j \mathbf{K}^{-1}\tilde{\mathbf{u}}_j$, where $\ell_j > 0$ is scalar depth, and $\tilde{\mathbf{u}}_j = [\mathbf{u}_j^\top, 1]^\top$ is the keypoint in homogeneous coordinates. By construction, sliding along the ray does not change the reprojection:

$$\pi_{\mathbf{K}}(\mathbf{X}_j) = \mathbf{u}_j, \qquad \frac{d}{d\ell_j}\pi_{\mathbf{K}}(\mathbf{X}_j) = \mathbf{0}. \tag{5}$$

The remaining problem is converting the resulting 3D keypoints to SMPL meshes, which is a learned inverse kinematics (IK) problem. In practice, not all 3D keypoint samples are plausible SMPL meshes. Empirically, however, we find that they are generally strong hypotheses such that the IK decoded mesh joints do not deviate far from them. Here the formulation applies only to the visible joints. Occluded or truncated joints have no direct image constraint, but remain informed by image context and anatomic dependencies between keypoints.

Probabilistic factorisation. We can now factorise the probability distribution over body states as follows:

$$\begin{aligned} p(\mathbf{z} \mid \mathbf{I}, \mathbf{K}) = \iint & p_{\psi}(\mathbf{z} \mid \mathbf{X}(\mathbf{u}, \boldsymbol{\ell}; \mathbf{K}), \mathbf{I}, \mathbf{K}) \\ & p_{\boldsymbol{\omega}_\ell}(\boldsymbol{\ell} \mid \mathbf{u}, \mathbf{I}, \mathbf{K}) p_{\boldsymbol{\omega}_u}(\mathbf{u} \mid \mathbf{I}, \mathbf{K})\, \mathrm{d}\boldsymbol{\ell}\, \mathrm{d}\mathbf{u}, \end{aligned} \tag{6}$$

where $\mathbf{u} \in \mathbb{R}^{J\times 2}$ denotes the 2D keypoints, $\boldsymbol{\ell} \in \mathbb{R}^J$ their depths along the rays from the camera centre to the keypoints, and $\mathbf{X}(\mathbf{u}, \boldsymbol{\ell}; \mathbf{K})$ the camera-frame 3D keypoints. As discussed in the previous paragraph, we take a three-stage approach:

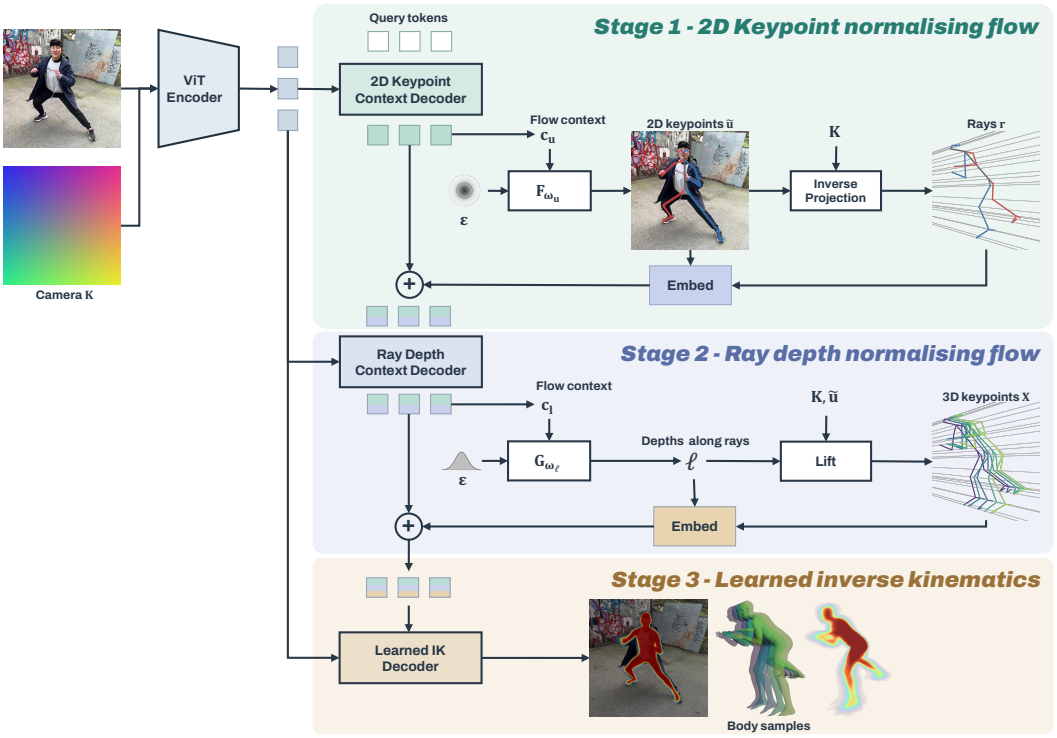

Figure 3: **Ray-Flow HMR model.** We encode the image and camera information using a vision transformer. Then, we decode a body sample in three stages. First, an image-space normalising flow draws 2D keypoint samples. These are embedded as additional conditioning for the next stage, a depth-along-the-ray normalising flow which predicts 3D keypoints as distances along camera rays. This explicitly enforces 3D keypoint consistency. Finally, a learned inverse-kinematics decoder maps each set of predictions to a body sample.

- Stage-1 draws 2D keypoint samples from $p_{\boldsymbol{\omega}_u}$.
- Stage-2 draws the corresponding depths from $p_{\boldsymbol{\omega}_\ell}$ conditioned on 2D results. These are then deterministically lifted to become 3D keypoint samples using the camera.
- Stage-3 applies a learned inverse-kinematics (IK) decoder to obtain pose, shape, and translation parameter samples.

We keep image and camera features available as additional conditioning for each stage. These provide useful information beyond the first stage: Distance along the ray needs information from the image in addition to 2D keypoints; inverse kinematics resolves joint twists and body plausibility from the image as well. Our model Ray-Flow HMR is illustrated in Figure 3. Next, we discuss each stage in more detail.

Stage-0: Ray-aware image encoding. We first encode the cropped image with a vision transformer (ViT) (Xu et al., 2022). To preserve the camera geometry discarded by cropping, we map every crop pixel back into the full image and backproject it through $\mathbf{K}^{-1}$ to construct a dense map of unit camera-ray directions. A convolution embeds this ray map into patch tokens, which are added to the corresponding image patch embeddings before the ViT blocks. The resulting features are therefore conditioned on both image appearance and the camera rays associated with the crop.

Stage-1: Image-space 2D keypoint distribution. The first sampling stage models the distribution of 2D keypoints $p_{\boldsymbol{\omega}_u}(\mathbf{u} \mid \mathbf{I}, \mathbf{K})$ with a conditional normalising flow. We construct the flow context with a transformer decoder comprising J learned keypoint queries. These go through self-attention layers to exchange information across keypoints and model anatomical dependencies, and cross-attention layers to attend to image evidence. We concatenate these tokens as conditioning context for

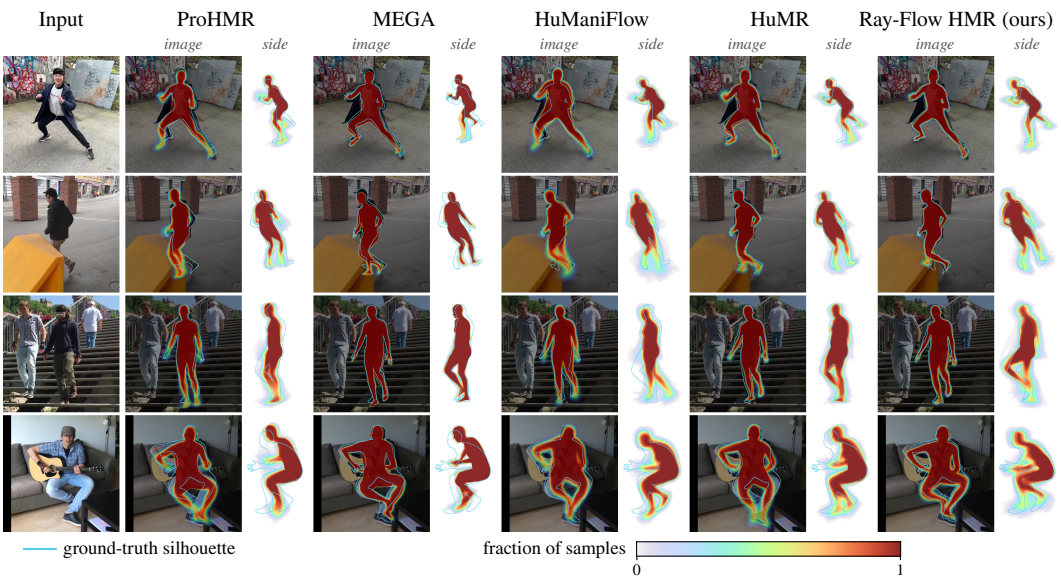

Figure 4: **Sample distributions on EMDB (top two rows) and 3DPW (bottom two rows).** For each method, 100 body samples are rendered as a heatmap of the fraction of samples covering each pixel, in the image and from a side view, with the ground-truth silhouette outlined in cyan. Front-view heatmaps show *consistency*: how well samples align with visible joints. A good model should be 'reddish'. Side-view heatmaps show *diversity*: how well monocular uncertainty is captured. A good model should be 'bluish'. Ray-Flow HMR heatmaps are more consistent and more diverse.

the normalising flow $F_{\boldsymbol{\omega}_u}$. Pushing Gaussian noise samples through the model gives us 2D keypoint samples

$$\boldsymbol{\epsilon}_u^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{2J}), \qquad \mathbf{u}^{(s)} = F_{\boldsymbol{\omega}_u}\left(\boldsymbol{\epsilon}_u^{(s)}; \mathbf{I}, \mathbf{K}\right). \tag{7}$$

Stage-2: Depth-along-the-ray distribution. Each sampled 2D keypoint $\mathbf{u}_j^{(s)}$ defines a camera ray through the optical centre. After mapping the keypoint back to full-image coordinates, inverse-projection through $\mathbf{K}^{-1}$ gives the ray direction. Conditioned on the sampled skeleton and image features, a second normalising flow predicts how far along each ray the corresponding joint lies:

$$\mathbf{X}_j^{(s,r)} = \ell_j^{(s,r)} \mathbf{K}^{-1} \tilde{\mathbf{u}}_j^{(s),\text{full}}. \tag{8}$$

The distances are sampled jointly to capture dependencies across the body. By construction, each resulting 3D keypoint lies on its corresponding ray and therefore projects back to the sampled 2D location.

Stage-3: Learned inverse kinematics. The final stage maps each lifted 3D keypoint set to SMPL pose and shape. The decoder also receives image features, which provide cues about body shape and local rotations that are not fully determined by the keypoints. We place the decoded body at the sampled root:

$$\begin{aligned} (\boldsymbol{\theta}^{(s,r)}, \boldsymbol{\beta}^{(s,r)}) &= D_\psi\left(\mathbf{X}^{(s,r)}, \mathbf{I}, \mathbf{K}\right), \\ \mathbf{t}^{(s,r)} &= \mathbf{X}_{\text{root}}^{(s,r)} - \mathcal{J}_{\text{root}}\left(\boldsymbol{\theta}^{(s,r)}, \boldsymbol{\beta}^{(s,r)}\right). \end{aligned} \tag{9}$$

The decoder is deterministic, so variation in the final meshes comes from the preceding flows. Because not every lifted keypoint set is exactly realisable by SMPL, the decoded joints may deviate from their rays. Empirically, we found that Stage-2 samples are sufficiently plausible such that the IK decoder does not have to deviate far from them.

3.3 TRAINING OBJECTIVES AND INFERENCE

We train Ray-Flow HMR with the following losses. We use different loss functions for synthetic batches and real batches. We empirically discover that training naively with real data improves accuracy and consistency, but significantly reduces sample diversity. We argue that this is likely due to the fact that real data labels are fitted themselves, and as such carry bias from the fitting priors.

Flow likelihoods. Given a synthetic batch, we minimise the negative log likelihood of the ground truth 2D joints $\mathbf{u}^\star$ for the Stage-1 normalising flow, as well as the NLL of the ground truth depth for the Stage-2 normalising flow

$$\mathcal{L}_{\text{2D}} = -\mathbb{E}[\log p_{\boldsymbol{\omega}_u}(\mathbf{u}^\star \mid \mathbf{I}, \mathbf{K})] \qquad \mathcal{L}_{\text{3D}} = -\mathbb{E}[\log p_{\boldsymbol{\omega}_\ell}(\boldsymbol{\ell}^\star \mid \mathbf{u}^\star, \mathbf{I}, \mathbf{K})]\,. \tag{10}$$

Intuitively, the loss should encourage the model to assign high likelihood to the ground truth joints. To maximise image conditional likelihood, the flow should be certain for visible joints, and be uncertain for occluded ones. We treat a real batch differently and supervise the zero-noise stage-1 flow sample with an L1 loss only $\mathcal{L}_{\text{2D,real}} = ||\hat{\mathbf{u}} - \mathbf{u}^*||_1$. We discover that this helps with generalisation without harming diversity. We detail this via an ablation in Appendix B.

Inverse-kinematics losses. We train the IK decoder with both the GT and the sampled keypoints. Given the GT, we supervise the decoded SMPL parameters, keypoints, vertices, and pose rotations via L1 losses. Given a sample, we minimise the distances between Stage-2 and Stage-3 3D keypoints:

$$\mathcal{L}_{\text{IK,sample}} = \mathbb{E}_s \left[\sum_{j \neq \text{root}} \left\| \mathcal{J}_j(\boldsymbol{\theta}^{(s)}, \boldsymbol{\beta}^{(s)}) + \mathbf{t}^{(s)} - \mathbf{X}_j^{(s)} \right\|_1 \right]. \tag{11}$$

The root is omitted because Equation 9 matches it exactly. We also regularise the twist of each sampled pose towards the decoder's prediction from the same image and its ground-truth skeleton. Each sample is therefore matched to its own lifted joints, rather than regressed to the single ground-truth body or its projection.

Inference. At inference, we sample $\mathbf{u}$, recompute its depth context, sample $\boldsymbol{\ell}$, lift the pair with Equation 8, and decode it with Equation 9. Repeating the depth draw for a fixed $\mathbf{u}$ exposes ambiguity along a fixed set of rays; drawing both variables produces samples from the full cascade. For a deterministic prediction, we pass zero base noise through both flows and then decode.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and baselines. We train our models using BEDLAM (Black et al., 2023) which provides exact body ground truths used to render the images. Additionally, we use 4DHumans (Goel et al., 2023) data to improve model generalisability. We evaluate without fine-tuning on the 3DPW (Von Marcard et al., 2018) test set and EMDB1 (Kaufmann et al., 2023).

We compare with the deterministic CameraHMR (Patel & Black, 2025) reference and four probabilistic methods: ProHMR (Kolotouros et al., 2021), HuManiFlow (Sengupta et al., 2023), MEGA (Fiche et al., 2025), and HuMR (Wehrbein et al., 2025). HuMR uses an HRNet-W48 (Sun et al., 2019) image backbone and ViTPose-H (Xu et al., 2022) detections; its released backbone is fine-tuned on 3DPW-train. More information on implementation details is given in Appendix A.

Metrics. We adopt the accuracy–consistency–diversity evaluation introduced by HuManiFlow. *Point accuracy* is evaluated using Procrustes-aligned mean per-joint position error (PA-MPJPE) of each method's native point prediction; for RayFlow-HMR, this is the sample generated from zero-noise flow input. We evaluate *sample accuracy* as the minimum PA-MPJPE among $S = 100$ draws for each image (Min.). *Consistency* is the mean Euclidean reprojection error to annotated 2D joints, in pixels in the common 256 crop. We report it for both the point estimate and all samples. On 3DPW it is evaluated on visible COCO-17 (Lin et al., 2014) joints; on EMDB it uses all annotated SMPL-24 joints. *Diversity* is the mean 3D distance of every sampled COCO-17 joint from that joint's sample

Table 1: Distribution quality metrics using $S = 100$ samples. Diversity uses the shared HuMR visibility rule. The horizontal rule separates ViT-Base and ViT-Huge models; best metrics are bolded. †: fine-tuned on 3DPW-train.

Method	3DPW (Von Marcard et al., 2018)			EMDB (Kaufmann et al., 2023)		
	PA-MPJPE (mm) ↓ Point / Min.	Consistency (px) ↓ Point / Sample	Diversity (mm) ↑ Vis. / Invis.	PA-MPJPE (mm) ↓ Point / Min.	Consistency (px) ↓ Point / Sample	Diversity (mm) ↑ Vis. / Invis.
ProHMR (Kolotouros et al., 2021)	60.05 / 48.26	7.95 / 8.70	34.77 / 47.18	87.44 / 70.50	9.21 / 9.83	38.37 / 54.08
MEGA (Fiche et al., 2025)	47.25 / 40.71	7.97 / 7.93	11.15 / 19.95	71.27 / 61.38	7.66 / 7.65	14.11 / 27.60
HuManiFlow (Sengupta et al., 2023)	54.30 / 40.65	5.09 / 6.39	42.82 / 70.51	81.98 / 59.87	8.25 / 9.41	53.72 / 95.84
Ours (base)	**44.31 / 31.95**	**3.97 / 4.51**	**46.68 / 87.27**	**59.55 / 41.02**	**4.89 / 5.78**	**57.28 / 108.12**
CameraHMR (Patel & Black, 2025)	40.89 / –	**3.53** / –	–	**46.61** / –	4.27 / –	–
HuMR (Wehrbein et al., 2025)	40.86 / 29.75†	4.33 / 5.29†	35.03 / **75.41**†	56.10 / 40.90	4.73 / 5.68	41.01 / **88.27**
Ours (huge)	**40.50 / 29.52**	3.76 / **4.18**	**40.43** / 66.92	51.33 / **37.09**	**4.21 / 4.80**	**45.08** / 79.34

mean, reported separately for visible and invisible joints using one frozen ViTPose-H confidence mask shared across methods. Following HuMR (Wehrbein et al., 2025), a joint is input-visible when its detector confidence exceeds 0.5 and input-invisible otherwise. All 3D errors and diversity values are in millimetres. All stochastic methods draw $S = 100$ samples.

4.2 Comparison with deterministic and probabilistic HMR

Quantitative results. Earlier baselines use ViT-Base backbones, whereas more recent methods scale up to ViT-Huge. Therefore, we train two variants of Ray-Flow HMR, the base model for baseline comparisons, and the huge model for generalisability and point estimation accuracy. Table 1 compares our models to existing baselines. The base model comprehensively improves minimum-sample accuracy, consistency, and diversity over ProHMR, MEGA, and HuManiFlow on both datasets. The huge model provides greater point accuracy and sample consistency. We outperform HuMR in terms of accuracy and consistency while retaining higher visible diversity on EMDB and 3DPW without fine-tuning; Figure 5 plots the rate of change in sample accuracy and sample consistency of our model vs the probabilistic baselines. Here, minimum PA-MPJPE rate of improvement indicates distribution accuracy; best and tail sample consistency indicate distribution alignment with the image constraints. Ray-Flow HMR is best in all three cases on the two evaluation datasets.

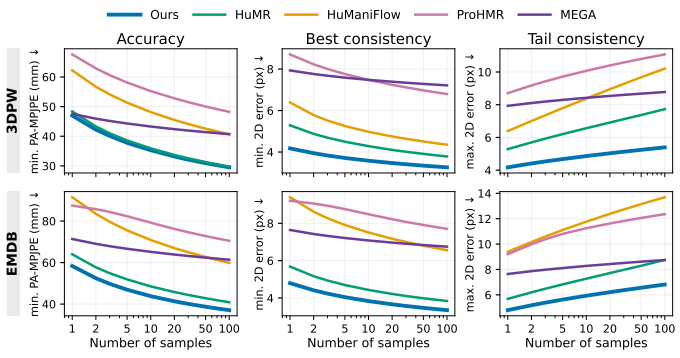

Figure 5: **Rate of change in minimum sample PA-MPJPE and minimum and maximum 2D consistency.** Faster accuracy improvements indicate a more accurate distribution; faster best-consistency improvements and slower tail-consistency degradation indicate better image alignment.

Qualitative results. Figure 4 qualitatively compares the predictive distributions of ProHMR, MEGA, HuManiFlow, and RayFlow-HMR on EMDB and 3DPW. For each method, front- and side-view heatmaps show the occupancy of 100 body samples, with the ground-truth silhouette outlined in cyan. The front view reveals whether samples are aligned with the image evidence, whereas the side view exposes variation along dimensions that are ambiguous from a monocular image. Intuitively, the plausible predictive distribution should be tight (or reddish) in the visible regions, while being diverse (or blue/greenish) in the occluded regions of the front view, as well as along the dimension exposed by the side view. ProHMR is moderately consistent and diverse but excels at neither; MEGA is concentrated in the front view but also narrow from the side; HuManiFlow and HuMR exhibit substantial side-view variation at the cost of greater front-view dispersion. In contrast, RayFlow-HMR

combines concentrated front-view support with broad, structured side-view variation, consistent with the quantitative results. We would like to highlight the tight sample silhouettes in the visible regions of our front-view renderings, which demonstrates the effectiveness of our ray constraints. We also show mesh samples from our model in Figure 1. Ray-Flow HMR generates consistent samples while remaining diverse in the ambiguous dimension. We provide more qualitative results in Appendix C.

Uncertainty-guided view selection. We conduct a simple experiment to show that our uncertainty is useful in practical applications such as multi-view reconstruction view selection. Here, we render synthetic humans from THuman2.1 (Yu et al., 2021), and feed the images into EasyMocap (eas, 2021) to reconstruct the body mesh. We test random selection, as well as uncertainty-guided selection using our model where we greedily add the candidate whose predicted uncertainty is lowest along the directions of greatest residual uncertainty under the views selected so far. As shown in Figure 6, our uncertainty-guided policy reduces MPJPE substantially faster. The improvement remains consistent at four and eight views, showing that the predicted uncertainty identifies complementary observations for multiview reconstruction. The shaded 95% confidence intervals are obtained by resampling and recomputing the mean error.

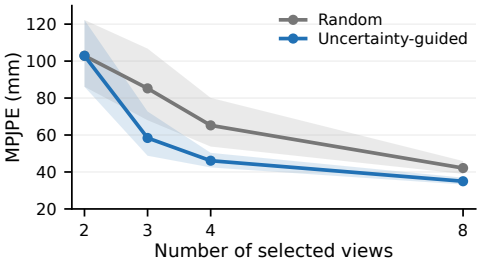

Figure 6: **View selection.** MPJPE over 100 subjects from a shared two-view seed; shading shows 95% bootstrap intervals. Uncertainty provides meaningful cues for view selection.

5 Discussion and Limitations

We do not claim state-of-the-art *point estimation accuracy* compared with existing scaling-law models such as SMPLer (Cai et al., 2023), SMPLest (Yin et al., 2025), and Sam-3D-Body (Yang et al., 2026) due to compute and data constraints. Instead, we show that our model is comparable to CameraHMR (Patel & Black, 2025) a strong deterministic baseline trained on similarly sized datasets.

Our learned distribution is only an approximation of the true data distribution. Although synthetic training datasets provide perfect ground truths, they are rendered into the environment with random poses without correlation with the background. In the real-data distribution, interactions of people and the environment provide additional semantic constraints beyond image evidence, which our model does not fully capture. In the ablation study, we identified the harmful biases carried by the fitted pseudo-labels on in-the-wild image data. Future work should include real data with robust 3D ground truth such as Harmony4D (Khirodkar et al., 2024) to incorporate proper real-data supervision into the model.

6 Conclusion

In this paper, we proposed Ray-Flow HMR, a probabilistic HMR method that generates consistent and diverse body samples. We motivated our work by analysing the complexity of the consistency manifold defined by keypoint reprojection alignment with image evidence. We argued that learning a distribution that samples from this manifold is difficult. Instead, we factorised the full predictive distribution into a few stages: a 2D keypoint distribution that solved the easier 2D problem, then a depth-along-the-ray distribution for 3D keypoint samples constrained by the 2D solutions, and finally an inverse kinematic stage that regressed the final body samples. We showed that this factorisation is effective in terms of learning a more plausible predictive distribution, with evaluations on accuracy–consistency–diversity metrics. We demonstrated that Ray-Flow HMR generates much more consistent samples than existing baselines, while being more diverse.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to retrieve related works, to coding experiments, and to prove mathematical claims. We have not used generative AI tools to generate synthetic datasets, or to draft sections of the paper. We have reviewed all AI-assisted work. We have checked citation correctness manually. We have reviewed AI generated code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Easymocap - make human motion capture easier. Github, 2021. URL `https://github.com/zju3dv/EasyMocap`.

Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in neural information processing systems*, 33:20496–20507, 2020.

Christopher M Bishop. Mixture density networks. 1994.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8726–8737. IEEE, 2023.

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*, pp. 561–578. Springer, 2016.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, and others. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *European Conference on Computer Vision*, pp. 146–165. Springer, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J Black, and Dimitrios Tzionas. Poco: 3d pose and shape estimation with confidence. In *2024 International Conference on 3D Vision (3DV)*, pp. 85–95. IEEE, 2024a.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1323–1333. IEEE, 2024b.

Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8781–8791. IEEE, 2023.

Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. MEGA: Masked Generative Autoencoder for Human Mesh Recovery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5366–5378. IEEE, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14737–14748. IEEE, 2023.

Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1381–1388. IEEE, 2009.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131. IEEE, 2018.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14586–14597. IEEE, 2023.

Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37:107270–107285, 2024.

Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11107–11117. IEEE, 2021.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2252–2261, 2019.

Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic Modeling for Human Mesh Recovery. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11585–11594. IEEE, 2021.

Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9879–9887. IEEE, 2019.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3382–3392. IEEE, 2021.

Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12933–12942. IEEE, 2023.

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 *European Conference on Computer Vision*, pp. 590–606. Springer, 2022.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1954–1963. IEEE, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2021.

George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22:57:1–57:64, 2021. URL https://jmlr.org/papers/v22/19-1028.html.

Priyanka Patel and Michael J Black. CameraHMR: Aligning People with Perspective. In *2025 International Conference on 3D Vision (3DV)*, pp. 1562–1571. IEEE, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pp. 10975–10985, 2019.

François Rozet et al. Zuko: Normalizing flows in pytorch, 2022. URL https://pypi.org/project/zuko.

Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Pu Wang, Hongfei Xue, Srijan Das, and Chen Chen. Genhmr: Generative human mesh recovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6749–6757, 2025.

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.

Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11199–11209. IEEE, 2021a.

Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16089–16099. IEEE, 2021b.

Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so (3) manifolds for human pose and shape distribution estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4779–4789. IEEE, 2023.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5686–5696. IEEE, 2019.

Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European conference on computer vision*, pp. 614–631. Springer, 2018.

Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini De Mello, and Michael Stengel. Blade: Single-view body mesh learning through accurate depth estimation. *arXiv preprint arXiv:2412.08640*, 2024.

Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14598–14608. IEEE, 2023.

Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic Monocular 3D Human Pose Estimation With Normalizing Flows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11179–11188. IEEE, 2021.

Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Utilizing Uncertainty in 2D Pose Detectors for Probabilistic 3D Human Mesh Recovery. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5852–5862. IEEE, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Xitong Yang, Devansh Kukreja, Don Pinkus, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jia-Wei Liu, Nicolás Ugrinovic, Anushka Sagar, and others. Sam 3d body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7209–7219, 2026.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, and others. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021.

Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13106–13115. IEEE, 2022.

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 11426–11436. IEEE, 2021.

A Implementation details

Training. We train ViT-Base and ViT-Huge variants of Ray-Flow HMR to provide a fair comparison with baselines. The models are initialised using ViT-Pose (Xu et al., 2022) weights. Images are cropped to the input size of 256×256 pixels. The conditional normalising flows are implemented using Zuko (Rozet et al., 2022). Both the 50-dimensional 2D keypoint flow and the conditional 3D keypoint flow contain eight affine coupling layers with two-layer, width-512 conditioners. Four samples per image were used for the sampled-IK objective during training.

We optimise with AdamW (Loshchilov & Hutter, 2017) using batch size 64, initial learning rate 10^{-5}, weight decay 10^{-4}, and a cosine schedule ending at 10^{-7} after 317,500 updates. Training is done on a single RTX6000 Blackwell GPU with bfloat16 mixed precision and takes roughly 1.5 days.

Evaluation. At inference, Ray-Flow HMR uses intrinsics estimated by FLNet (Patel & Black, 2025) on 3DPW and the released intrinsics on EMDB. For comparison, every method's projected joints are mapped into the same 256-pixel reporting crop before computing metrics such as consistency of the 2D reprojected keypoints in pixel space. We use mesh-regressed H36M/LSP-14 joints on 3DPW and all 24 SMPL joints on EMDB. For EMDB, we define the pelvis as the midpoint of the two hips and fit the Procrustes transform over the same joints that are scored.

B Bias in real-image datasets

Real data provide important generalisability and accuracy improvements to our model. However, it is difficult to extract the right information from them. We argue this is mainly because SMPL labels for in-the-wild images are generated by optimisation with terms such as reprojection and differentiable silhouette losses, together with several priors. This yields precise image alignment but introduces bias that a probabilistic model should avoid, particularly for fitted out-of-frame or self-occluded joints and joint depths that are not explicitly constrained by image evidence.

Table 2: Exploratory real-data supervision ablation on EMDB1 using ViT-H and $S = 100$. Diversity uses the shared HuMR visibility rule.

Real-image supervision	PA-MPJPE (mm) ↓ Point / Min.	Consistency (px) ↓ Point / Sample	Diversity (mm) ↑ Vis. / Invis.
Full real-data supervision	46.59 / 35.86	3.96 / 4.32	32.74 / 58.47
UV NLL + sampled-IK ray loss	50.91 / 37.81	4.31 / 4.68	41.34 / 69.02
UV NLL only	51.26 / 37.88	4.43 / 4.84	43.36 / 71.40
UV point L_1 only (adopted)	51.33 / 37.09	4.21 / 4.80	45.08 / 79.34
None (synthetic only)	50.87 / 37.09	4.49 / 5.23	51.84 / 98.05

In Table 2, we ablate different real-image supervision strategies. Full supervision improves accuracy and consistency, but substantially reduces sample diversity. We progressively remove depth supervision for Stage 2 and parameter losses for Stage 3, then replace the Stage 1 NLL with an L_1 loss on the zero-noise sample. The final configuration provides the best trade-off between preserving sample diversity and retaining gains in accuracy and consistency.

C Additional Qualitative Results

We provide some more qualitative results in the next few pages. We also attach, in the supplementary, a simple webpage to provide closer inspection of these results.

We show additional results on 3DPW in Figure 7.

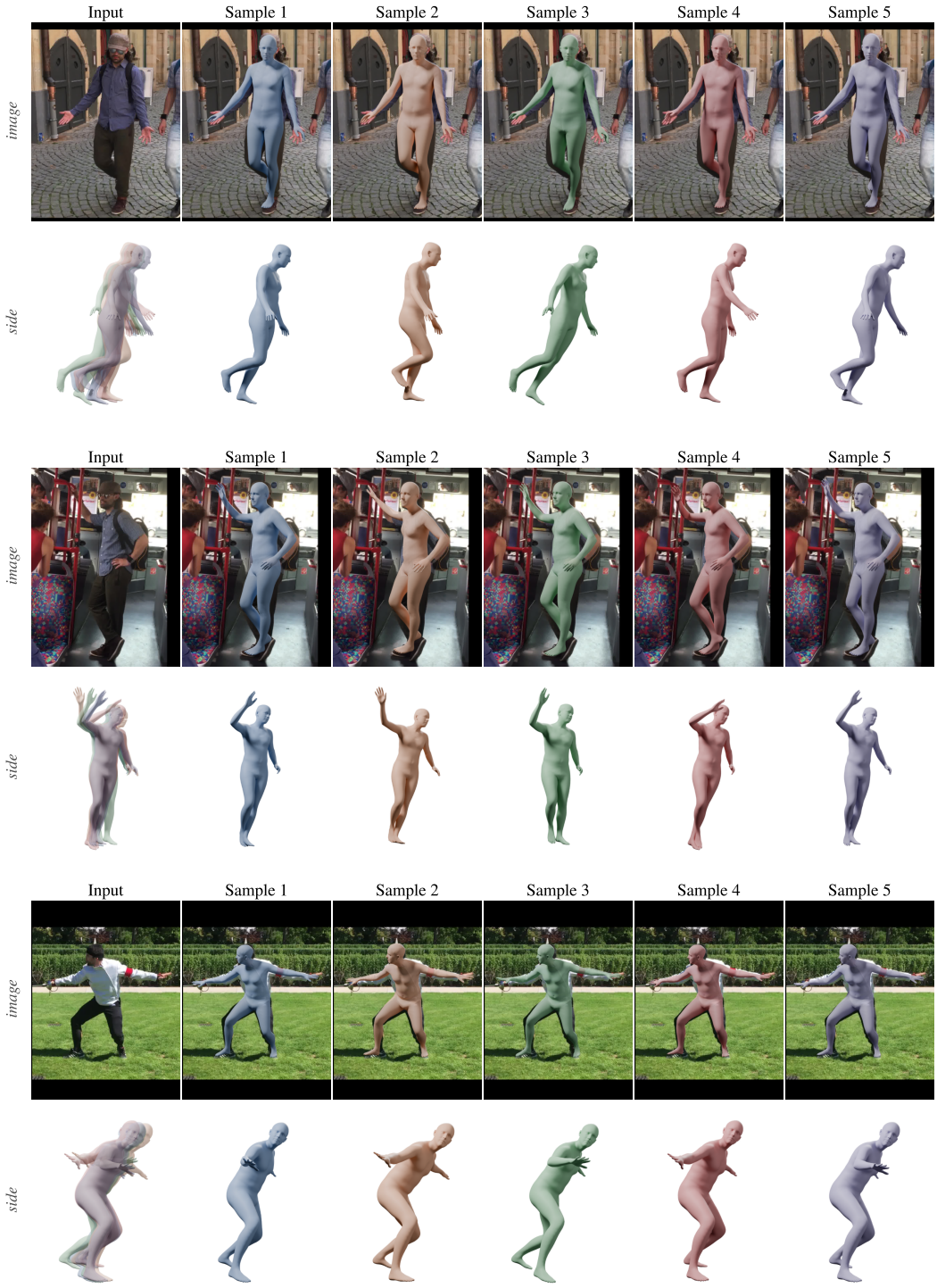

Figure 7: Additional Ray-Flow HMR samples on 3DPW.

We show additional results on EMDB in Figure 8.

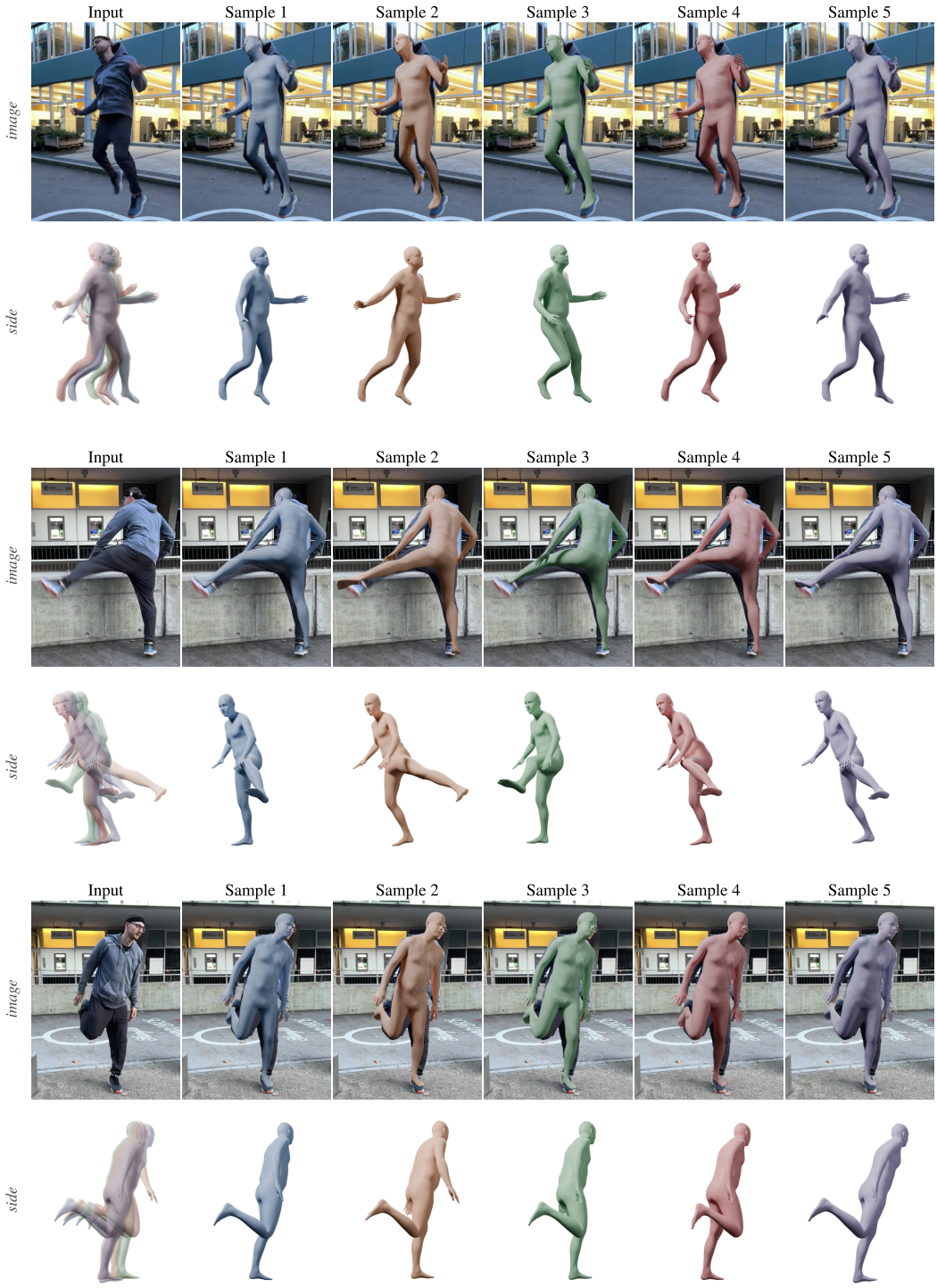

Figure 8: Additional Ray-Flow HMR samples on EMDB.

We show additional sample heatmap results in Figure 9.

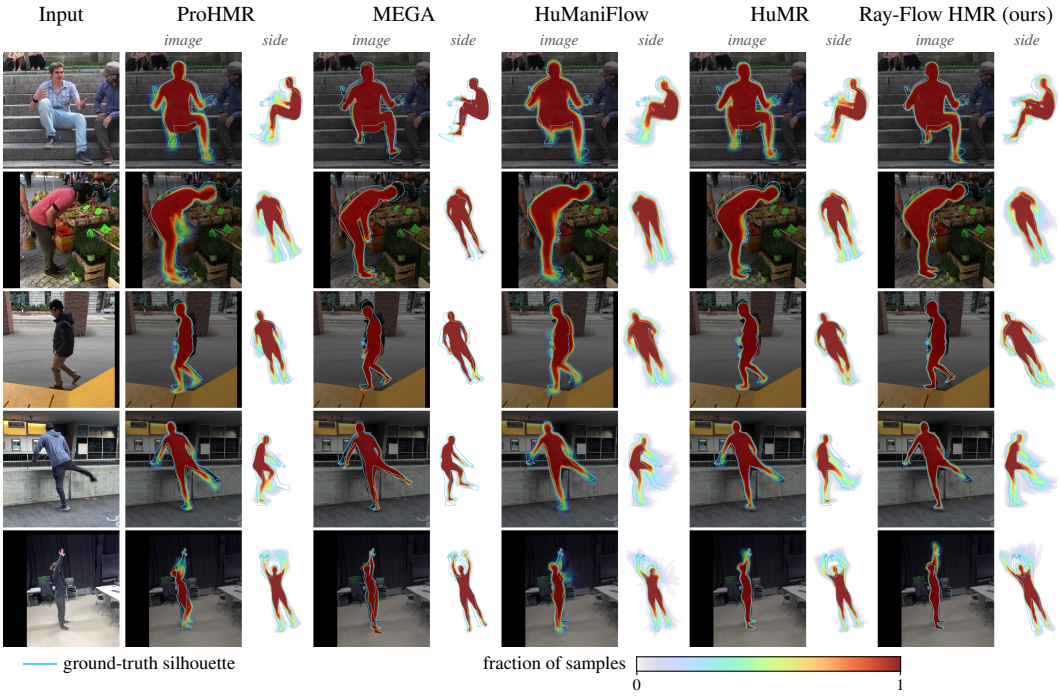

Figure 9: Additional predictive-distribution comparisons on 3DPW and EMDB.

Failure cases. We show two failure cases for our model in Figure 10. Row 1 is failure caused by lack of generalisability to a difficult pose. Row 2 is arguably a *semantic* failure: The person is holding her right ankle with her hand, but model samples do not capture this. Samples do not always put the right hand close to the right ankle.

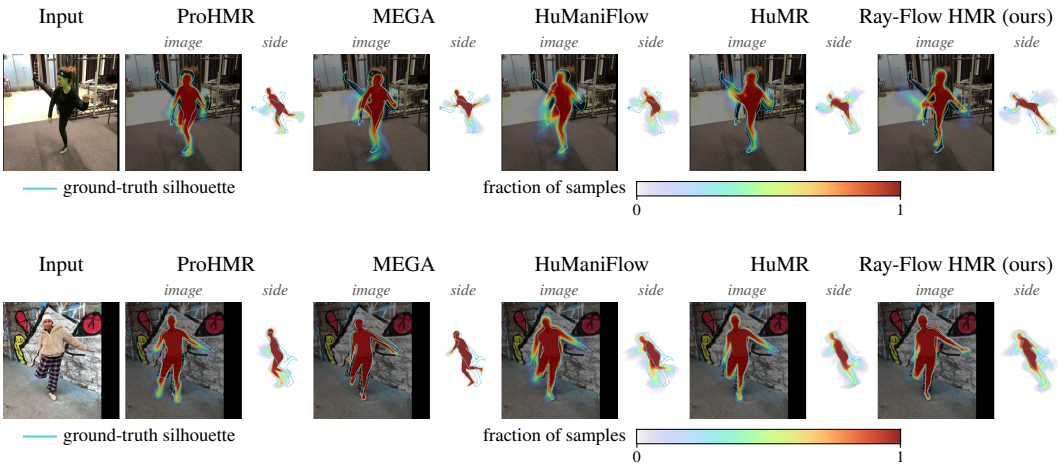

Figure 10: Failure cases. Row 1 fails due to difficulty of the pose; Row 2 fails to capture the semantics in the image - the person is holding her ankle with her hand.