PIXELBODY: PIXEL-ALIGNED HUMAN MESH RECOVERY FROM MONOCULAR VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward human mesh recovery predicts 3D bodies from images and video in one shot but does not *fit* them to the input pixels, so the results can disagree with image cues such as landmarks, silhouettes, and surface normals. PixelBody is an offline method that starts from the body trajectory of a feed-forward method and iteratively fits it to pixel-aligned cues. During the fitting process, it utilizes a pose autoencoder pretrained on motion capture data. However, instead of applying it as a regularization, PixelBody adapts the decoder to each person: its weights are first fine-tuned to reproduce the initial trajectory from the encoded pose latents, and then the decoder and latents are fitted together to the video cues. Our ablations show that fine-tuning the decoder prevents the fit from being limited to the motion-capture prior, while still benefiting from it, and achieves lower 3D error than both direct pose optimization and a frozen decoder. Fitting runs in two stages: registration to skeleton and surface landmarks, then refinement that adds silhouettes, body normals, and relative depth. The landmarks, normals, and depth come from predictors we fine-tuned from Sapiens2 and Wan 2.1. PixelBody has the lowest camera-space error on EMDB1 among published methods that recover a body mesh, and, with ground-truth cameras, the lowest world-space error on EMDB2 among the compared methods.

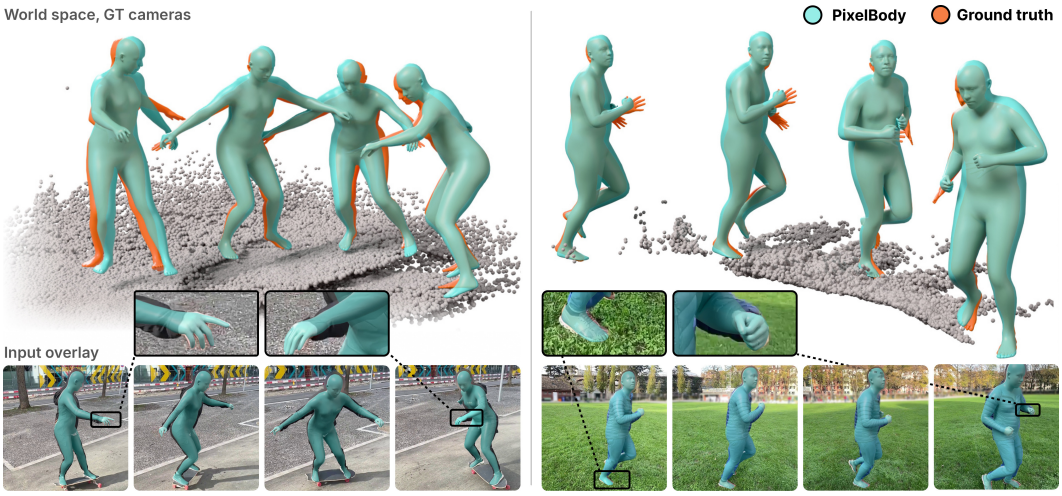

Figure 1: PixelBody predictions (teal) and ground truth (orange) on two EMDB clips, visualized in world space using ground-truth cameras. The input frames alongside the reconstructed mesh overlays are provided below. Gray points represent a COLMAP-recovered scene that is provided as a reference. Note that the ground truth in this dataset has no finger pose.

1 INTRODUCTION

Monocular human mesh recovery (HMR) aims to estimate 3D human pose, shape, and motion from a single RGB video. Feed-forward models estimate these quantities from RGB in one shot (Goel et al.,

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.

2023; Patel & Black, 2025; Wang et al., 2025b; Li et al., 2025; Yang et al., 2026b), but their estimates are not explicitly optimized for dense alignment with the input pixels, leaving discrepancies between the recovered body and the observed person.

Dense iterative fitting can align bodies with per-pixel cues, such as surface correspondences, silhouettes, normals, or depth. A number of methods have primarily applied such dense fitting to single frames of monocular images, point clouds, or calibrated multi-view captures (Xiu et al., 2022; 2023; Zhang et al., 2026b; Cai et al., 2026; Wang et al., 2026b). On the other hand, monocular *video* fitting is typically limited to silhouettes (Xu et al., 2018; Alldieck et al., 2018), sparse keypoints (Rempe et al., 2021; Ye et al., 2023) or surface landmarks (Hewitt et al., 2024; Cuevas-Velasquez et al., 2026). These methods optimize pose or pose latents, relying on a frozen prior that pushes their results towards its training distribution, potentially limiting the expressivity of their results.

Some HMR methods go beyond their pre-trained motion priors by using test-time adaptation. CycleAdapt (Nam et al., 2023) adapts a regressor and a motion denoiser over a test video, and STRIDE (Lal et al., 2025) fine-tunes a pretrained motion prior on each video's 3D estimates. NeMo (Wang et al., 2023a) fits a neural motion field to HMR estimates of several videos of one action and then optimizes the field's weights and instance codes jointly. None of these methods adapt an existing motion prior to fit per-pixel cues in the video.

Figure 1 shows results of PixelBody, our offline method that combines dense image cues with test-time adaptation of the motion model. Since the image feature quality is crucial to obtain the best results, we first trained Sapiens2 Body and Wan2.1 Body, which are finetuned versions of Sapiens2 (Khirodkar et al., 2026) and Wan 2.1 (Wan Team, 2025). These predictions serve as reference targets for our optimization process. Inspired by pivotal tuning (Roich et al., 2023), we encode the initial SAM 3D Body (Yang et al., 2026b) estimations with a frozen encoder and fine-tune the decoder to reproduce the trajectory from those latents. Then, our two-stage PixelSolver optimizes pose latents and decoder weights to skeleton and surface landmarks, followed by refining the results using silhouettes, normals, and relative depth.

Our goal with PixelBody is to create an HMR system that works for in-the-wild data. In that spirit, we do not train on the evaluation datasets, including their training subsets. We primarily evaluate our approach on EMDB (Kaufmann et al., 2023) and RICH (Huang et al., 2022), measuring 3D body error in camera and world coordinates. On EMDB, PixelBody has the lowest camera-space error among published mesh-recovery methods and improves on its SAM 3D Body initialization on all three metrics. With ground-truth cameras, it also has the lowest world-space error on the EMDB2 subset among the compared methods.

Thus, our contributions are:

- **Pose decoder adaptation:** a pretrained pose decoder and its latents are fitted to explain the motion of each person in the image, reducing the 3D error compared to direct pose optimization.
- **PixelSolver:** a whole-video solver that registers the trajectory to landmarks and then refines it against silhouettes, normals, and relative depth, with identity, bone lengths, and focal length shared over the sequence.
- **PixelBody:** PixelSolver with our Sapiens2 Body and Wan2.1 Body predictors, with the lowest camera-space EMDB error among published mesh-recovery methods. We plan to release the solver, the AMASS-pretrained pose decoder, and our Sapiens2-based predictors.

2 Related Work

2.1 Image, Video, and World-Grounded HMR

Single-image HMR has moved from direct regression (Kanazawa et al., 2018) to fitting in the training loop (Kolotouros et al., 2019), reprojection feedback (Zhang et al., 2021; Wang & Daniilidis, 2023), diffusion-based optimization (Stathopoulos et al., 2024; Ho et al., 2025; Lu et al., 2025), camera cues (Kocabas et al., 2021; Li et al., 2022; Wang et al., 2023b; Patel & Black, 2025), and recent feed-forward systems (Goel et al., 2023; Sárándi & Pons-Moll, 2024; Wang et al., 2025b; Fiche et al., 2026; Yang et al., 2026b). Video HMR uses temporal encoders, learned motion priors,

or diffusion (Kanazawa et al., 2019; Kocabas et al., 2020; Choi et al., 2021; Zheng et al., 2025; Li et al., 2025; Qian et al., 2026). World-grounded systems recover human motion in world coordinates (Sun et al., 2023; Shin et al., 2024; Wang et al., 2024; Shen et al., 2024; Yin et al., 2024; Yang et al., 2025; Zhao et al., 2026; Yang et al., 2026a). DuoMo lifts camera-space diffusion estimates to world coordinates with estimated camera motion and refines them with world-space diffusion (Wang et al., 2026d). Others reconstruct people and scenes (Liu et al., 2026; Song et al., 2026; Chen et al., 2026; Li et al., 2026a; Shi et al., 2026; Sur et al., 2026). PixelBody does not reconstruct the scene, so comparisons with them use body error only. SAM-Body4D and WorldMotion (Gao et al., 2025; Tu et al., 2025) smooth tracked SAM 3D Body outputs, the latter in world coordinates, whereas PixelBody fits them to pixel-aligned predictions in camera space.

2.2 Pixel-Aligned Geometry and Dense Body Fitting

DenseReg, DensePose, and continuous surface embeddings established dense image-to-surface correspondence (Güler et al., 2017; 2018; Neverova et al., 2020), later used for mesh reconstruction and fitting (Güler & Kokkinos, 2019; Yao et al., 2022; Lê et al., 2024). DenseRaC, DSR, and PEAR train body regressors with render-and-compare or pixel-level losses (Xu et al., 2019; Dwivedi et al., 2021; Wu et al., 2026).

DenseSMPLify fits bodies to Sapiens-predicted normal maps of single images (Zhang et al., 2026b; Khirodkar et al., 2024). ICON and ECON fit bodies to predicted normals and silhouettes of clothed humans (Xiu et al., 2022; 2023), and other single-image methods to predicted depth (Yalandur Muralidhar et al., 2025; Bravo-Sánchez et al., 2026). OmniFit targets 3D surface landmarks predicted from possibly scale-distorted point clouds (Cai et al., 2026). DiffProxy uses dense correspondences predicted from calibrated multi-view images by a diffusion model trained on synthetic data (Wang et al., 2026b). Earlier monocular video methods fit a personalized template (Xu et al., 2018) or a body model with offsets (Alldieck et al., 2018) to joint detections and silhouettes. Hewitt et al. (2024) and MAMMA (Cuevas-Velasquez et al., 2026) fit bodies to predicted surface landmarks in monocular or multi-view video, and CameraHMR does so per image to build pseudo ground truth (Patel & Black, 2025). PixelBody instead fits a body trajectory over a monocular video to Sapiens2 Body landmarks and, by rendering, to silhouettes and Wan2.1 Body normals and depth.

Image and video diffusion models are repurposed for depth and normals (Ke et al., 2024; He et al., 2025; Hu et al., 2025; Bin et al., 2025; He et al., 2026), and human-centric models predict them from images (Khirodkar et al., 2024; Saleh et al., 2025) or video (Kim et al., 2025; Wang et al., 2026a; Miao et al., 2026). Wan2.1 Body is also a video diffusion model fine-tuned on synthetic data, but it predicts normals and depth of the body model in its template shape rather than of the clothed person, so they can be compared with body renders.

2.3 Sequence Fitting and Test-Time Adaptation

Classical methods fit bodies to image cues under learned pose priors (Bogo et al., 2016; Pavlakos et al., 2019). Video fitting adds learned motion models and typically keeps them fixed: HuMoR optimizes latent transitions of a pretrained conditional variational autoencoder (VAE) (Rempe et al., 2021), SLAHMR and PACE couple such priors with camera optimization (Ye et al., 2023; Kocabas et al., 2024), GLAMR infills occluded motion before global refinement (Yuan et al., 2022), and Yu et al. (2026) learn the prior as neural distance fields. Diffusion motion models are used through denoising (Zhang et al., 2024), noise optimization (Karunratanakul et al., 2024; Lyu et al., 2026), score distillation (Li et al., 2024), or surface-landmark guidance (Wang et al., 2026d). HTD-Refine fits HMR outputs to predicted temporal dynamics (Wei et al., 2026).

Test-time adaptation instead updates learned models (Tung et al., 2017). EFT, BOA, and DynaBOA tune image regressors per example or online (Joo et al., 2021; Guan et al., 2021; 2023), and Nie et al. (2024) meta-train a regressor as the initialization for per-image tuning. CycleAdapt (Nam et al., 2023) alternately adapts a regressor and a motion denoiser on a test video, Wen et al. (2026) extend it online to long single-person video streams with codebook anchor motions, and Lin et al. (2025) guide such adaptation with action labels from a vision-language model through a text-aligned motion prior. STRIDE fine-tunes a pretrained motion prior at test time on each video's 3D estimates (Lal et al., 2025). NeMo fits a neural motion field to HMR estimates of several videos of one action and then optimizes the field's weights and instance codes jointly (Wang et al., 2023a), whereas

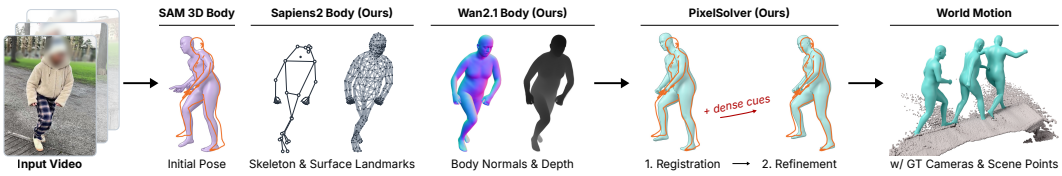

Figure 2: PixelBody overview. SAM 3D Body (Yang et al., 2026b) initializes the trajectory. Then, PixelSolver registers it to Sapiens2 Body landmarks and refines it with silhouettes and Wan2.1 Body normals and depth while adapting the pose decoder (Figure 3).

PixelBody's decoder is pretrained on motion capture data. Test-time training and deep image priors likewise update weights per test input (Sun et al., 2020; Ulyanov et al., 2018).

PixelBody first fits its pose decoder to the feed-forward trajectory with the encoded latents fixed, in the spirit of pivotal tuning (Roich et al., 2023), and then optimizes the decoder and the latents against video landmarks and dense cues, as DGP, IAGAN, and Re-Depth Anything do for images (Pan et al., 2020; Abu Hussein et al., 2020; Bhattarai & Rhodin, 2026). For multi-view HMR, Xie et al. (2026) find decoder updates less effective than optimizing learned SMPL tokens, while we observe that adapting a motion-VAE decoder to one video significantly outperforms optimization of the latents.

3 Method

3.1 Overview

Figure 2 outlines the components of PixelBody. The input is a monocular RGB video with person tracks, given as per-frame boxes and SAM 3 (Carion et al., 2026) masks. Three networks provide per-frame results to PixelSolver: SAM 3D Body (Yang et al., 2026b) predicts the initial body estimate, Sapiens2 Body predicts skeleton and surface landmarks with confidences, and Wan2.1 Body predicts camera-space body normals and relative depth. We fine-tune Sapiens2 Body from Sapiens2 (Khirodkar et al., 2026) and Wan2.1 Body from Wan 2.1 (Wan Team, 2025), as described in Appendices B.1 and B.2. PixelSolver first fits a per-person copy of a pretrained pose decoder to the initial trajectory. Next, two stages jointly optimize the decoders and latents over all people and frames. Landmark registration runs without rendering, and dense refinement then adds rendered silhouette, normal, and depth losses.

We model each person using the Momentum Human Rig (MHR; Ferguson et al., 2025). Specifically, we estimate per-frame root rotation and translation in camera coordinates, along with body and hand articulation coefficients. Additionally, body-shape identity coefficients and bone-length coefficients are shared across all frames. We assume that all frames in one sequence share one focal length and that the principal point is the full-frame center. We denote SAM 3D Body's feed-forward body-articulation trajectory by $\boldsymbol{\theta}_{\text{init}}$ and PixelSolver's trajectories by $\widehat{\boldsymbol{\theta}}$. Identity and bone length start from the medians of the SAM 3D Body estimates, and focal length from the median MoGe-2 (Wang et al., 2025a) estimate. Before each fitting stage, each frame's root is rescaled to this shared camera and body. Appendix A.1 details this rescaling and a root-depth repair for re-acquired 3DPW tracks.

3.2 Pose Decoder Adaptation

PixelSolver generates each person's body articulation with the pose decoder of a motion VAE and adapts the decoder and its pose latents, as Figure 3 shows. The VAE uses the transformer architecture of MLD (Chen et al., 2023) and encodes each N-frame window of body articulation into nine 256-dimensional latent tokens. For each person, R pose windows with frame sets $\mathcal{W}_r$ cover every track segment of at least N frames. Consecutive windows of a segment share an overlap of K frames, with values in Appendix A.1. The frozen encoder $\mathcal{E}$ first maps each window of $\boldsymbol{\theta}_{\text{init}}$ to pose latents $\mathcal{Z}_r$, which are then decoded by $\mathcal{D}$. Finally, overlap blending $\mathcal{B}$ averages the decoded windows into the trajectory $\widehat{\boldsymbol{\theta}}$:

$$\mathcal{Z}_r \leftarrow \mathcal{E}(\boldsymbol{\theta}_{\text{init}}[\mathcal{W}_r])\,, \qquad \widehat{\boldsymbol{\theta}} = \mathcal{B}(\mathcal{D}(\mathcal{Z}_1), \ldots, \mathcal{D}(\mathcal{Z}_R))\,. \tag{1}$$

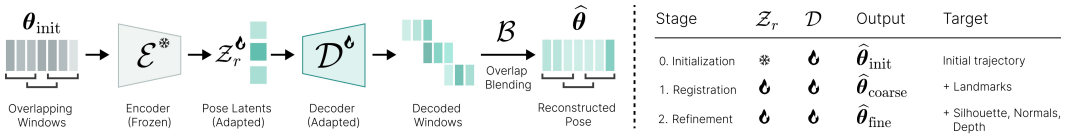

Figure 3: PixelSolver overview. Left: the frozen encoder $\mathcal{E}$ maps windows of $\boldsymbol{\theta}_{\text{init}}$ once to pose latents $\mathcal{Z}_r$, which the decoder $\mathcal{D}$ and overlap blending $\mathcal{B}$ turn into $\widehat{\boldsymbol{\theta}}$. Right: we showcase, per stage, whether $\mathcal{Z}_r$ and $\mathcal{D}$ are frozen (snowflake) or adapted (flame), as well as the reconstruction targets.

The initialization stage fits each person's copy of the decoder $\mathcal{D}$, shared across that person's windows, with $\mathcal{Z}_r$ fixed, so that $\widehat{\boldsymbol{\theta}}_{\text{init}}$ closely reproduces $\boldsymbol{\theta}_{\text{init}}$ before the fitting starts. During registration, we then optimize $\mathcal{Z}_r$ and $\mathcal{D}$ jointly and produce $\widehat{\boldsymbol{\theta}}_{\text{coarse}}$. Finally, the refinement carries the registered latents and decoder forward without re-encoding and outputs $\widehat{\boldsymbol{\theta}}_{\text{fine}}$. We provide more details in the Appendix A.1.

No loss ties $\mathcal{D}$ to its pretrained weights or $\mathcal{Z}_r$ to the VAE prior; the regularization and smoothness terms of Section 3.4 are applied directly to the body parameters. The pretrained decoder thus initializes and parameterizes the optimization, without limiting the motion space.

3.3 Pixel-Aligned Targets and Renders

Wan2.1 Body predicts normals and depth on N'-frame windows, each of which is one solver chunk, and consecutive chunks share a K'-sized overlap with values in Appendix A.1. Unlike pose windows, which follow each person's track segments, solver chunks are defined on the video and shared by all people. Within a chunk, dense targets, masks, and renders share one crop, as described in Appendix A.1. DRTK (Pidhorskyi et al., 2024) renders each chunk at half the crop resolution with one z-buffer across people for normals and depth and a separate silhouette per person.

Each person's occluder mask marks the pixels outside the tracking mask that MoGe-2 depth places in front of that person. The person's silhouette is the tracking mask minus other people's tracking masks and the occluder mask. It is further restricted to pixels with valid predicted normals, to the person's box, and to the crop. Landmarks outside the dilated tracking mask or inside the occluder mask receive zero confidence. Because Wan2.1 Body is trained on synthetic bodies with zero identity, all renders use the template shape with the current bone lengths. The silhouette, normal, and depth losses therefore update pose and bone length but not identity. Surface landmarks use the current identity, so their reprojection loss updates it.

3.4 Fitting Objective and Optimization

Both fitting stages optimize hand articulation, root transforms, identity, and bone length, plus $\mathcal{Z}_r$ and $\mathcal{D}$ within pose windows. Refinement also optimizes focal length and a sequence-wide depth scale and offset. Each fitting stage minimizes

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{smooth}}. \tag{2}$$

The data term $\mathcal{L}_{\text{data}}$ contains confidence-weighted reprojection losses on the Sapiens2 Body landmarks, not on ground-truth 2D annotations, with hand skeleton landmarks up-weighted. Refinement adds a silhouette loss inside each person's box, excluding other people's silhouettes. It also adds normal and depth losses on the union of all silhouettes. The normal loss ignores pixels of body parts whose skeleton landmarks all have confidence at most 0.25. Predicted depth is normalized per chunk. After registration, one robust least-squares solve (Ranftl et al., 2022) fixes the chunks' relative scales and offsets over the sequence based on 1) rendered depth from the registered trajectory and 2) neighboring chunks' depth on shared pixels.

The regularizer $\mathcal{L}_{\text{reg}}$ contains a zero-pose term, MHR parameter limits, and two anchors. An early-registration anchor keeps low-confidence skeleton landmarks near their feed-forward positions. A trajectory anchor pulls $\widehat{\boldsymbol{\theta}}$ toward $\boldsymbol{\theta}_{\text{init}}$ in registration and toward $\widehat{\boldsymbol{\theta}}_{\text{coarse}}$ in refinement. In registration, the smoothness term $\mathcal{L}_{\text{smooth}}$ penalizes finite differences of root and joint transforms. In refinement, it penalizes deviations of local and global (camera-space) joint positions and rotations

Table 1: Camera-space errors (mm) on EMDB1 and RICH. Italic: ground-truth intrinsics. ‡: 3D joints only, no mesh. See Appendix Table 6 for the full comparison and 3DPW metrics. Dark/medium/light green: first/second/third; gray: unranked.

	EMDB1 (24)			RICH (24)		
Method	**PA-MPJPE↓**	**MPJPE↓**	**PVE↓**	**PA-MPJPE↓**	**MPJPE↓**	**PVE↓**
Image-based						
NLF-L+fit, RICH-trained (Sárándi & Pons-Moll, 2024)	40.9	68.4	80.6	28.7	51.0	58.2
PromptHMR (Wang et al., 2025b)	*41.0*	*71.7*	*84.5*	*37.3*	*56.6*	*65.5*
SAM 3D Body (Yang et al., 2026b)	38.2	61.7	72.5	30.9	53.7	60.3
Video-based						
GVHMR (Shen et al., 2024)	42.7	72.6	84.2	*39.5*	*66.0*	*74.4*
TRAM (Wang et al., 2024)	*45.7*	*74.4*	*86.6*	–	–	–
GENMO (Li et al., 2025)	42.5	73.0	84.8	*39.1*	*66.8*	*75.4*
DuoMo (Wang et al., 2026d)	*41.8*	*67.1*	*78.2*	*34.8*	***51.4***	***58.9***
Human3R (Chen et al., 2026)	48.5	73.9	86.0	–	–	–
OnlineHMR (Zhao et al., 2026)	46.0	74.0	86.1	–	–	–
THFM‡ (Wang et al., 2026a)	35.2	54.6	–	34.7	55.0	–
SHOW (Shi et al., 2026)	45.8	75.7	88.8	–	–	–
Ours (Sapiens2 Body, AMASS decoder)	35.8	61.3	73.3	33.4	59.4	67.1
Ours (Wan2.1 Body, AMASS decoder)	**35.1**	59.1	70.8	32.2	57.4	65.0
Ours (Full Model)	35.2	**59.0**	**70.5**	**30.6**	54.8	61.7

Table 2: World space with ground-truth cameras on EMDB2 (mm, 100-frame segments, 24 joints; RTE: % of root path). Dark/medium/light green: first/second/third; RTE is not ranked.

Method	**WA-MPJPE$_{100}$↓**	**W-MPJPE$_{100}$↓**	**RTE↓**
TRAM (Wang et al., 2024)	62.7	182.8	0.2
GENMO (Li et al., 2025)	69.5	185.9	0.9
PromptHMR-vid (Wang et al., 2025b)	52.9	142.1	0.2
DuoMo, our run (Wang et al., 2026d)	52.8	136.1	0.2
MetricHMSR (Song et al., 2026)	55.6	152.5	0.2
SAM 3D Body, our run (Yang et al., 2026b)	51.5	134.7	0.2
Ours (Sapiens2 Body, AMASS decoder)	46.1	125.4	0.2
Ours (Wan2.1 Body, AMASS decoder)	45.3	124.6	0.2
Ours (Full Model)	**44.7**	**122.6**	0.2

from their 7-frame Savitzky–Golay-filtered versions (Savitzky & Golay, 1964). Registration runs for 600 AdamW iterations and refinement for 500. Appendix A details the terms, weights, and schedules.

4 Experiments

4.1 Evaluation Protocol

We evaluate PixelBody on EMDB (Kaufmann et al., 2023) and RICH (Huang et al., 2022), and on 3DPW (von Marcard et al., 2018) in Appendix Table 6. Tables 1 and 2 report PixelBody with Sapiens2 Body and the pose decoder trained on AMASS (Mahmood et al., 2019), where a Sapiens2 Body normal and depth predictor replaces only Wan2.1 Body's normal and depth targets. We also evaluate a version that uses Wan2.1 Body with the same AMASS-trained decoder, which the ablations and figures also use. Finally we evaluate our full model, which uses a pose VAE of the same architecture pretrained on a larger private motion-capture corpus, see Appendix A for more details. EMDB1 and EMDB2 are the official 17- and 25-sequence EMDB subsets, and EMDB1+2 is their union.

We convert MHR to SMPL (Loper et al., 2015) by optimization with the official MHR toolkit and report PA-MPJPE and MPJPE on 24 SMPL joints and PVE over all SMPL vertices in camera space (Yang et al., 2026b), and WA-MPJPE$_{100}$, W-MPJPE$_{100}$, and RTE in world space (Shin et al., 2024), all in millimeters except RTE, which is in percent of the ground-truth root path length. Appendices A.1, C.2, and C.1 give the solver settings, the protocol, and the inputs.

4.2 Camera-Space Evaluation

Table 1 reports each method under its own input protocol; PixelBody uses ground-truth-derived crops on EMDB1 and ground-truth-derived boxes on RICH, with track prompts on both. On

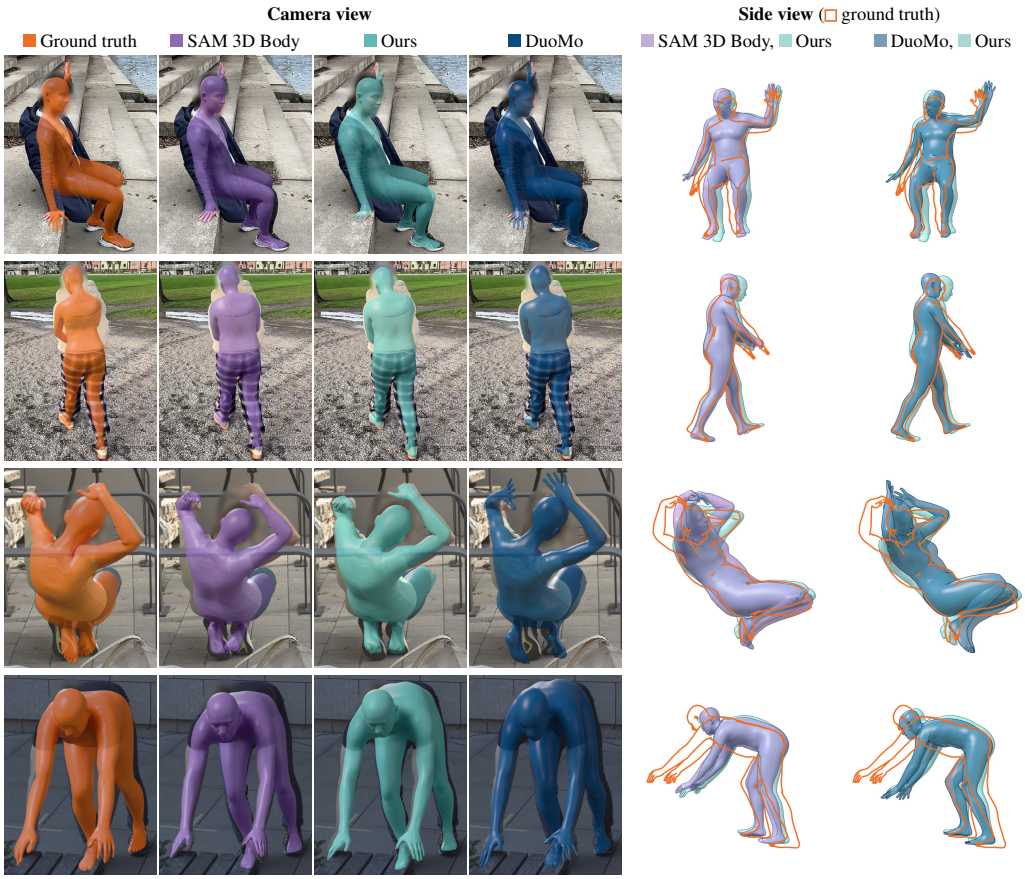

Figure 4: Two EMDB1 and two RICH clips in camera space: ground truth, SAM 3D Body, PixelBody, and DuoMo (with ground-truth intrinsics, here also given the ground-truth body height, unlike in our tables), and pelvis-aligned side views with the ground truth outlined; faces are blurred.

EMDB1, PixelBody has the lowest error on all three metrics among published mesh-recovery methods with the AMASS-decoder model and with the Full Model; THFM, which predicts joints but no mesh, has lower MPJPE. Our fitting procedure lowers the PA-MPJPE of the mask-box SAM 3D Body initializer of Table 3a, which shares shape and focal length within each sequence, by 4.2 mm on EMDB1+2.

On RICH, PixelBody reaches 33.4 mm PA-MPJPE with Sapiens2 Body, 32.2 mm with the AMASS-decoder model, and 30.6 mm with the Full Model, against 30.9 for SAM 3D Body and 28.7 for NLF-L+fit, which is trained on RICH. In Appendices C, D.1, and E.2, we discuss the 3DPW and RICH results in more detail. Figure 4 compares PixelBody with its SAM 3D Body initialization and with DuoMo, the strongest video-based mesh-recovery baseline, on two EMDB1 and two RICH clips. SAM 3D Body and PixelBody use their estimated focal lengths, and in the side views the other method is drawn over PixelBody.

We choose frames for the visualization from the ground truth alone: each clip gets a different ground-truth pose class, upright, crouching, bending, or lying, and shows its most distinctive frame of that class.

4.3 Temporal and World-Space Accuracy

In Table 3a, we report that fitting lowers Jerk from 9.9 at PixelSolver's starting point to 3.6, near the ground truth's 3.0. Smoothing the starting point alone with a Savitzky-Golay filter reaches 2.7 but changes its PA-MPJPE by only about 0.5 mm, against about 4 mm for fitting, so we read Jerk as a sanity check, not the entire evaluation.

Table 3: Solver ablations of the AMASS-decoder model (mm; PA: PA-MPJPE; WA, W: WA-/W-MPJPE$_{100}$): (a) pipeline; (b) decoder vs. direct poses; (c) adapted blocks; (d) dense cues. Dark/medium/light green: first/second/third, (b) per pair, Jerk nearest GT; (a), RTE: unranked.

		EMDB1+2				EMDB2, GT cameras		
Stage	**Variant**	**PA**	**MPJPE**	**PVE**	**Jerk**	**WA**	**W**	**RTE**
(a) Solver stages: pipeline								
SAM 3D Body	Mask boxes	36.1	56.5	69.0	10.5	51.5	134.7	0.20
	GT boxes	33.3	53.4	64.3	7.9	–	–	–
Starting point	Mask boxes	36.1	56.7	69.3	9.9	50.8	134.0	0.20
Smoothing only	Mask boxes	35.7	56.2	68.6	2.7	49.9	132.3	0.20
AMASS-decoder model		31.9	52.4	64.1	3.6	45.3	124.6	0.18
Ground truth		–	–	–	3.0	–	–	–
(b) Solver stages: decoder vs. direct, matched								
Registration	Direct	35.2	56.5	68.2	**4.7**	49.9	135.1	0.18
	Decoder	**33.3**	**55.2**	**67.5**	5.6	**48.4**	**131.7**	0.20
Refinement	Direct	33.5	53.8	65.1	**3.0**	46.6	126.4	0.17
	Decoder	**31.9**	**52.4**	**64.1**	3.6	**45.3**	**124.6**	0.18
(c) Adapted blocks								
Decoder		**31.8**	**52.1**	**63.7**	**3.6**	**45.1**	**123.6**	0.18
Latents		40.6	62.7	74.6	5.3	54.1	145.5	0.19
Both (AMASS decoder)		31.9	52.4	64.1	**3.6**	45.3	124.6	0.18
(d) Dense cues								
No dense cues		32.7	54.9	67.0	3.7	47.6	128.6	0.20
+ Normals		**31.8**	52.4	64.2	**3.6**	45.6	126.0	0.19
+ Normals + depth		**31.8**	**52.3**	**64.0**	**3.6**	45.5	124.6	0.18
+ Normals + silhouette		31.9	**52.3**	**64.0**	**3.6**	**45.3**	**123.2**	0.19
All three (AMASS decoder)		31.9	52.4	64.1	**3.6**	**45.3**	124.6	0.18

PixelBody does not model world motion. In Table 2, with ground-truth cameras on EMDB2, PixelBody has lower WA-MPJPE$_{100}$ and W-MPJPE$_{100}$ than every compared method with the AMASS-decoder model and with the Full Model, 45.3 and 124.6 mm and 44.7 and 122.6 mm, against 51.5 and 134.7 mm for our SAM 3D Body run lifted with the same cameras. We report estimated-camera results in Appendix Table 8. On RICH, DuoMo's WA-MPJPE$_{100}$ and W-MPJPE$_{100}$ of 53.6 and 80.8 mm compare with 52.5 and 81.6 mm for the AMASS-decoder model and 49.7 and 77.7 mm for the Full Model, Appendix Table 8. Figure 5 shows world-space overlays with DuoMo and JOSH, and Appendix G depicts visualizations after each training stage and the failure cases.

4.4 Ablations and Predictors

Table 3b compares adapting the decoder with optimizing the poses directly: the Decoder rows are the AMASS-decoder model's registration and refinement outputs, and the Direct rows come from one run that differs from it only in optimizing poses directly, Appendix E.1. Adaptation lowers error in both stages, PA-MPJPE by 1.9 mm after registration and 1.5 mm after refinement, at a cost in smoothness. In the pipeline rows, fitting lowers the error of the SAM 3D Body initialization with mask boxes on every metric. In Table 3c, adapting the decoder outperforms optimizing the latents of the frozen pretrained decoder, and adapting both, as PixelSolver does, matches adapting the decoder alone. In Table 3d, normals give most of the gain over the control without dense cues, and adding depth or silhouettes to them changes geometry little.

Finally, we evaluate the Sapiens2 Body landmark estimator. On ground-truth crops, checkpoints trained after the one PixelSolver uses have higher skeleton PCKh@0.1 than public pose predictors, 53.0% on EMDB and 51.0% on RICH against 46.9% and 42.9% of the closest baseline (Khirodkar et al., 2026), and, through a comparison head, lower surface-landmark error than MammaNet (Cuevas-Velasquez et al., 2026), 7.60 against 8.96 pixels on EMDB, see Appendix E.3. As the dense predictor in an earlier solver configuration, Wan2.1 Body gives 31.7 mm PA-MPJPE on EMDB1+2, against 33.0 mm for a Sapiens2 Body dense predictor, for more results see Appendix E.5. PixelBody's profiled components take about 2.9 s per frame, 3.4 s with Wan2.1 Body normals, excluding tracking and depth prediction, see Appendix F for a detailed analysis.

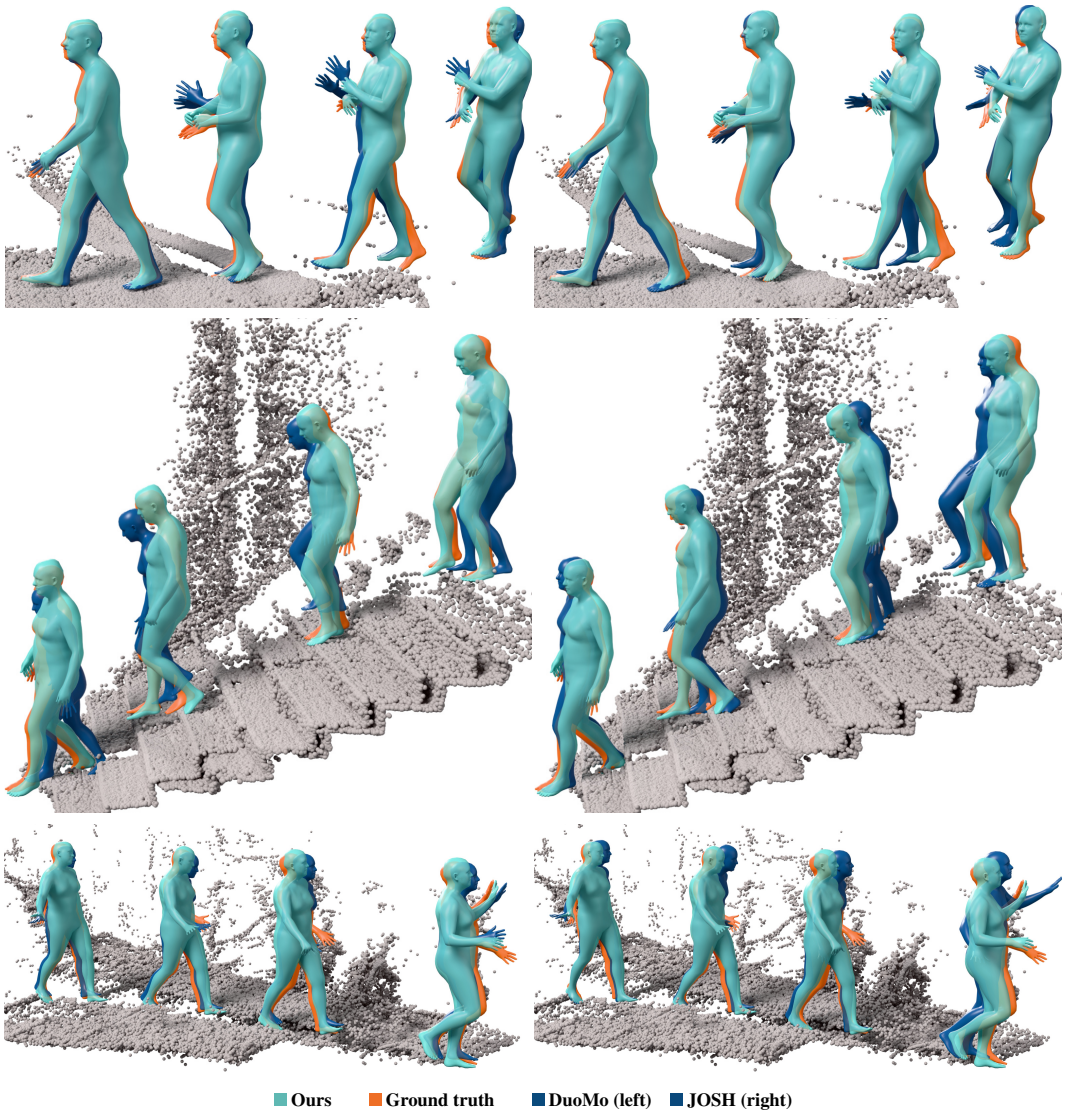

Figure 5: Three EMDB2 clips in world space: PixelBody and ground truth with DuoMo (left; given the ground-truth height as in its released code, not in our tables) and JOSH (right; released code, no 2D-keypoint term), placed with ground-truth cameras and one similarity per window.

5 Conclusion

We presented PixelBody, which fits feed-forward HMR trajectories to pixel-aligned predictions and silhouettes while adapting a pretrained pose decoder and its latents per person. On EMDB, it has the lowest camera-space error among published mesh-recovery methods. Under matched fitting settings, adapting the decoder lowers 3D error over direct pose optimization in both stages. Its evaluation crops, boxes, and track prompts derive from ground truth. Its limitations include reliance on the initializer and pre-computed person tracks. Our method also does not estimate camera motion; with ground-truth cameras, its world-space errors are the lowest among the compared methods. Extending the method to natively handle camera reconstruction remains future work.

AI use statement

We used generative AI tools in essentially every part of this work, including method design, experiment design, code implementation, running experiments, data processing, result analysis, literature

search, figures, and writing. The authors directed the research, reviewed all AI-assisted work, tested the code, and verified numbers, citations, and claims against experiment records and sources. The authors take full responsibility for the paper and the code.

Reproducibility statement

The appendix provides solver hyperparameters and loss weights for each stage. Our evaluation converts MHR to SMPL by optimization, with one SMPL identity per tracked person. We plan to release the PixelSolver inference code, the AMASS-pretrained pose decoder, and our Sapiens2-based landmark, normal, and depth predictors upon acceptance.

Ethics statement

We evaluate on public human-motion benchmarks and blur faces in all displayed frames. The training and evaluation data may not represent all bodies, clothing, environments, or motions, so performance should not be assumed to transfer uniformly across people. PixelBody is intended for human-motion research, not for identification, surveillance, or high-stakes decisions.

References

Shady Abu Hussein, Tom Tirer, and Raja Giryes. Image-adaptive GAN based reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 3121–3129, 2020.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8387–8397, 2018.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686–3693, 2014.

Ananta R. Bhattarai and Helge Rhodin. Re-Depth Anything: Test-time depth refinement via self-supervised re-lighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings (CVPR Findings)*, pp. 602–612, 2026.

Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. NormalCrafter: Learning temporally consistent normals from video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8330–8339, 2025.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, 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 (ECCV)*, pp. 561–578, 2016.

Laura Bravo-Sánchez, Matthieu Armando, Romain Brégier, Grégory Rogez, Serena Yeung-Levy, and Fabien Baradel. All-age human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3677–3687, 2026.

Zeyu Cai, Yuliang Xiu, Renke Wang, Zhijing Shao, Xiaoben Li, Siyuan Yu, Chao Xu, Yang Liu, Baigui Sun, Jian Yang, and Zhenyu Zhang. OmniFit: Multi-modal 3D body fitting via scale-agnostic dense landmark prediction. In *European Conference on Computer Vision (ECCV)*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and

Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, 2023.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *International Conference on Learning Representations (ICLR)*, 2026.

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1973, 2021.

Hanz Cuevas-Velasquez, Anastasios Yiannakidis, Soyong Shin, Giorgio Becherini, Markus Höschle, Joachim Tesch, Taylor Obersat, Tsvetelina Alexiadis, Eni Halilaj, and Michael J. Black. MAMMA: Markerless accurate multi-person motion acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7175–7186, 2026.

Alec Dewulf, Dhruv Pai, Li Yang, Ashley Zhang, and Ben Keigwin. Aurora: A leverage-aware spectral optimizer. *arXiv preprint arXiv:2606.27715*, 2026.

Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11250–11259, 2021.

Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, Igor Santesteban, Javier Romero, Jenna Zarate, Jeongseok Lee, Jinhyung Park, Jinlong Yang, John Doublestein, Kishore Venkateshan, Kris Kitani, Ladislav Kavan, Marco Dal Farra, Matthew Hu, Matthew Cioffi, Michael Fabris, Michael Ranieri, Mohammad Modarres, Petr Kadlecek, Rawal Khirodkar, Rinat Abdrashitov, Romain Prévost, Roman Rajbhandari, Ronald Mallet, Russell Pearsall, Sandy Kao, Sanjeev Kumar, Scott Parrish, Shoou-I Yu, Shunsuke Saito, Takaaki Shiratori, Te-Li Wang, Tony Tung, Yichen Xu, Yuan Dong, Yuhua Chen, Yuanlu Xu, Yuting Ye, and Zhongshi Jiang. MHR: Momentum Human Rig. *arXiv preprint arXiv:2511.15586*, 2025.

Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-HMR 2: Multi-person camera-centric human detection, mesh recovery and tracking. In *European Conference on Computer Vision (ECCV)*, pp. 281–299, 2026.

Mingqi Gao, Yunqi Miao, and Jungong Han. SAM-Body4D: Training-free 4D human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 753–762, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14783–14794, 2023.

Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10472–10481, 2021.

Shanyan Guan, Jingwei Xu, Michelle Zhang He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5070–5086, 2023.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. DepthFM: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 3203–3211, 2025.

Rıza Alp Güler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10884–10894, 2019.

Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. DenseReg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6799–6808, 2017.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7297–7306, 2018.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Xuanhua He, Jiaxin Xie, Mingzhe Zheng, and Qifeng Chen. ICDepth: Taming video diffusion models for video depth estimation via in-context conditioning. In *European Conference on Computer Vision (ECCV)*, pp. 657–674, 2026.

Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J. Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look Ma, no markers: Holistic performance capture without the hassle. *ACM Transactions on Graphics*, 43(6):1–12, 2024.

Hsuan-I Ho, Chen Guo, Po-Chen Wu, Ivan Shugurov, Chengcheng Tang, Abhay Mittal, Sizhe An, Manuel Kaufmann, and Linguang Zhang. PHD: Personalized 3D human body fitting with point diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7526–7537, 2025.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2015, 2025.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13274–13285, 2022.

Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pp. 42–52, 2021.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

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

Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5614–5623, 2019.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1334–1345, 2024.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14632–14643, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Zhaoen Su, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision (ECCV)*, pp. 206–228, 2024.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *International Conference on Learning Representations (ICLR)*, 2026.

Gwanghyun Kim, Xueting Li, Ye Yuan, Koki Nagano, Tianye Li, Jan Kautz, Se Young Chun, and Umar Iqbal. GeoMan: Temporally consistent human geometry estimation using image-to-video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7451–7461, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5253–5263, 2020.

Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11035–11045, 2021.

Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and camera motion estimation from in-the-wild videos. In *International Conference on 3D Vision (3DV)*, pp. 397–408, 2024.

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 (ICCV)*, pp. 2252–2261, 2019.

Rohit Lal, Saketh Bachu, Yash Garg, Arindam Dutta, Calvin-Khang Ta, Hannah Dela Cruz, Dripta S. Raychaudhuri, M. Salman Asif, and Amit K. Roy-Chowdhury. STRIDE: Single-video based temporally continuous occlusion-robust 3D pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 794–803, 2025.

Eric-Tuan Lê, Antonis Kakolyris, Petros Koutras, Himmy Tam, Efstratios Skordos, George Papandreou, Rıza Alp Güler, and Iasonas Kokkinos. MeshPose: Unifying DensePose and 3D body mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2405–2414, 2024.

Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. COIN: Control-inpainting diffusion prior for human and camera motion estimation. In *European Conference on Computer Vision (ECCV)*, pp. 426–446, 2024.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A GENeralist model for human MOtion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11766–11776, 2025.

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, Yuan Liu, and Yike Guo. UniSH: Unifying scene and human reconstruction in a feed-forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14038–14049, 2026a.

Moyang Li, Zihan Zhu, Marc Pollefeys, and Daniel Barath. DROID-SLAM in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36498–36508, 2026b.

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 (ECCV)*, pp. 590–606, 2022.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Qiuxia Lin, Rongyu Chen, Kerui Gu, and Angela Yao. Semantics-aware test-time adaptation for 3D human pose estimation. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 37780–37796, 2025.

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 (ECCV)*, pp. 740–755, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, and Ziwei Liu. DPoser-X: Diffusion model as robust 3D whole-body human pose prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9988–9997, 2025.

Xiaozhong Lyu, Korrawe Karunratanakul, Kaifeng Zhao, and Siyu Tang. LaMP: Learning robust latent motion prior for optimization-based human motion generation. In *International Conference on 3D Vision (3DV)*, pp. 893–903, 2026.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5442–5451, 2019.

Xingyu Miao, Junting Dong, Qin Zhao, Yuhang Yang, Junhao Chen, and Yang Long. From frames to sequences: Temporally consistent human-centric dense prediction. *arXiv preprint arXiv:2602.01661*, 2026.

Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3D human mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14829–14839, 2023.

Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 17258–17270, 2020.

Yongwei Nie, Mingxian Fan, Chengjiang Long, Qing Zhang, Jian Zhu, and Xuemiao Xu. Incorporating test-time optimization into training with dual networks for human mesh recovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 11634–11663, 2024.

Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *European Conference on Computer Vision (ECCV)*, pp. 262–277, 2020.

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

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Stanislav Pidhorskyi, Tomas Simon, Gabriel Schwartz, He Wen, Yaser Sheikh, and Jason Saragih. Rasterized edge gradients: Handling discontinuities differentiably. In *European Conference on Computer Vision (ECCV)*, pp. 335–352, 2024.

Zhiyin Qian, Siwei Zhang, Bharat Lal Bhatnagar, Federica Bogo, and Siyu Tang. Masked modeling for human motion recovery under occlusions. In *International Conference on 3D Vision (3DV)*, pp. 717–727, 2026.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11488–11499, 2021.

Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1–13, 2023.

Fatemeh Saleh, Sadegh Aliakbarian, Charlie Hewitt, Lohit Petikam, Xiao-Xian, Antonio Criminisi, Thomas J. Cashman, and Tadas Baltrušaitis. DAViD: Data-efficient and accurate vision models from synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5348–5358, 2025.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 140032–140065, 2024.

Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Boao Shi, Qiao Feng, Yiming Huang, and Lingjie Liu. Scene and human in one world: Reconstruction in a feedforward pass. *arXiv preprint arXiv:2606.27720*, 2026.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Chentao Song, He Zhang, Haolei Yuan, Haozhe Lin, Jianhua Tao, Hongwen Zhang, and Tao Yu. MetricHMSR: Metric human mesh and scene recovery from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21132–21142, 2026.

Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3D human recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 906–915, 2024.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 9229–9248, 2020.

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8856–8866, 2023.

Tanuj Sur, Shashank Tripathi, Nikos Athanasiou, Ha Linh Nguyen, Kai Xu, Michael J. Black, and Angela Yao. UniCon3R: Unified contact-aware 4D human-scene reconstruction from monocular video. *arXiv preprint arXiv:2604.19923*, 2026.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 54350–54382, 2025.

Zhangzheng Tu, Kailun Su, Shaolong Zhu, and Yukun Zheng. World-coordinate human motion retargeting via SAM 3D Body. *arXiv preprint arXiv:2512.21573*, 2025.

Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5236–5246, 2017.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9446–9454, 2018.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

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 (ECCV)*, pp. 614–631, 2018.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Kuan-Chieh Wang, Zhenzhen Weng, Maria Xenochristou, João Pedro Araújo, Jeffrey Gu, C. Karen Liu, and Serena Yeung. NeMo: 3D neural motion fields from multiple video instances of the same action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22129–22138, 2023a.

Letian Wang, Andrei Zanfir, Eduard Gabriel Bazavan, Misha Andriluka, and Cristian Sminchisescu. THFM: A unified video foundation model for 4D human perception and beyond. *arXiv preprint arXiv:2603.25892*, 2026a.

Renke Wang, Zhenyu Zhang, Ying Tai, Jun Li, and Jian Yang. DiffProxy: Multi-view human mesh recovery via diffusion-generated dense proxies. In *European Conference on Computer Vision (ECCV)*, 2026b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 35928–35959, 2025a.

Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3925–3935, 2023b.

Wenjia Wang, Liang Pan, Huaijin Pi, Yuke Lou, Xuqian Ren, Yifan Wu, Zhouyingcheng Liao, Lei Yang, Rishabh Dabral, Christian Theobalt, and Taku Komura. EmbodMocap: In-the-wild 4D human-scene reconstruction for embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28424–28434, 2026c.

Yufu Wang and Kostas Daniilidis. ReFit: Recurrent fitting network for 3D human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14644–14654, 2023.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *European Conference on Computer Vision (ECCV)*, pp. 467–487, 2024.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1148–1159, 2025b.

Yufu Wang, Evonne Ng, Soyong Shin, Rawal Khirodkar, Yuan Dong, Zhaoen Su, Jinhyung Park, Kris Kitani, Alexander Richard, Fabian Prada, and Michael Zollhöfer. DuoMo: Dual motion diffusion for world-space human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42777–42788, 2026d.

Dingkun Wei, Zehong Shen, Yan Xia, Georgios Pavlakos, Yujun Shen, and Xiaowei Zhou. Natural human motion recovery by aligning high-order temporal dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7187–7196, 2026.

Yilin Wen, Kechuan Dong, and Yusuke Sugano. Robust long-term test-time adaptation for 3D human pose estimation through motion discretization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 10612–10620, 2026.

Jiahao Wu, Yunfei Liu, Lijian Lin, Ye Zhu, Lei Zhu, Jingyi Li, and Yu Li. PEAR: Pixel-aligned Expressive humAn mesh Recovery. In *ACM SIGGRAPH 2026 Conference Papers*, pp. 1–12, 2026.

Haoyu Xie, Shengkai Xu, Cheng Guo, Muhammad Usama Saleem, Wenhan Wu, Chen Chen, Ahmed Helmy, Pu Wang, and Hongfei Xue. Monocular models are strong learners for multi-view human mesh recovery. In *European Conference on Computer Vision (ECCV)*, 2026.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13296–13306, 2022.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 512–523, 2023.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations (ICLR)*, 2025.

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. MonoPerfCap: Human performance capture from monocular video. *ACM Transactions on Graphics*, 37(2):1–15, 2018.

Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7760–7770, 2019.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 38571–38584, 2022.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 1212–1230, 2024.

Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. PhySIC: Physically plausible 3D human-scene interaction and contact from a single image. In *SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Fengyuan Yang, Kerui Gu, Ha Linh Nguyen, Tze Ho Elden Tse, and Angela Yao. Humans as checkerboards: Calibrating camera motion scale for world-coordinate human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6069–6079, 2025.

Fengyuan Yang, Tanuj Sur, Tze Ho Elden Tse, and Angela Yao. HumanBA: Human-aware bundle adjustment via global human-camera decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13846–13855, 2026a.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7209–7219, 2026b.

Chun-Han Yao, Jimei Yang, Duygu Ceylan, Yi Zhou, Yang Zhou, and Ming-Hsuan Yang. Learning visibility for robust dense human body estimation. In *European Conference on Computer Vision (ECCV)*, pp. 412–428, 2022.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21222–21232, 2023.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Ziwei Liu, and Lei Yang. WHAC: World-grounded humans and cameras. In *European Conference on Computer Vision (ECCV)*, pp. 20–37, 2024.

Zhengdi Yu, Simone Foti, Linguang Zhang, Amy Zhao, Cem Keskin, Stefanos Zafeiriou, and Tolga Birdal. Geometric neural distance fields for learning human motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2232–2242, 2026.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11038–11049, 2022.

Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Jing He, Zixin Zhang, Haodong Li, Yihao Liang, Kanghao Chen, Bin Ren, Xu Zheng, Shuai Yang, Kun Zhou, Yinchuan Li, Nicu Sebe, and Ying-Cong Chen. DVD: Deterministic video depth estimation with generative priors. *arXiv preprint arXiv:2603.12250*, 2026a.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11446–11456, 2021.

Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. RoHM: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14606–14617, 2024.

Zhaobo Zhang, Heming Li, Yongwei Nie, Chuhua Xian, and Guiqing Li. DenseSMPLify: 3D human body parametric reconstruction using pixel aligned dense normal maps. *Computer Graphics Forum*, pp. e70506, 2026b.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations (ICLR)*, 2025.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and László A. Jeni. OnlineHMR: Video-based online world-grounded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13951–13961, 2026.

Ce Zheng, Xianpeng Liu, Qucheng Peng, Tianfu Wu, Pu Wang, and Chen Chen. DiffMesh: A motion-aware diffusion framework for human mesh recovery from videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4891–4901, 2025.

Terms. Throughout the appendix, PixelSolver denotes PixelBody's two-stage solver, and the AMASS-decoder model denotes PixelBody with Wan2.1 Body, the AMASS decoder, and the PixelSolver configuration of Appendix A.1; the Full Model instead uses the decoder pretrained on a larger corpus, Appendix A. In pose decoder adaptation, PixelSolver fits a per-person copy of the pretrained pose decoder of a motion variational autoencoder (VAE), together with its pose latents, to the input video instead of holding them fixed as a motion prior. An initialization stage first fits each person's pose decoder with the pose latents fixed, so that the decoded trajectory reproduces the feed-forward trajectory. The two fitting stages, landmark registration and dense refinement, then optimize the pose latents and the pose decoder jointly.

In all tables, dark, medium, and light green mark the first, second, and third result per column, or per matched pair only the better one, lowest error, highest PCKh and FPS, or Jerk nearest the ground truth's 3.0; ties share a shade, and gray cells and RTE are unranked.

Appendix overview. Appendix A details PixelSolver: its inputs and initialization, pose decoder adaptation, the terms of the fitting objective, and the optimization schedule. Appendix B describes how the Sapiens2 Body and Wan2.1 Body predictors are trained and run. Appendix C details the benchmark inputs, metrics, evaluators, and sources of the comparison results. Appendix D lists all compared methods in camera and world space and reports PixelBody's performance on public benchmarks. Appendix E adds ablations for the solver, the initializer, the landmark predictors, and the dense predictors, with the confidence intervals behind the main paper's ablation claims. Appendix F reports runtime, and Appendix G adds qualitative results and failure cases.

A PixelSolver Details

A.1 Inputs, Objective, and Schedule

Per frame and tracked person, PixelSolver reads:

- MoGe-2 (Wang et al., 2025a) intrinsics and metric depth;
- SAM 3 (Carion et al., 2026) tracking masks and derived prompt boxes;
- SAM 3D Body (Yang et al., 2026b) MHR (Ferguson et al., 2025) pose, identity, bone lengths;
- Sapiens2 Body skeleton and surface landmarks with confidences;
- Wan2.1 Body (or Sapiens2 Body) normal and depth maps in chunk crops.

Dense targets, masks, and renders share the crop and cropped camera of their solver chunk, one 25-frame Wan2.1 Body window; chunks start 16 frames apart and share 9, with an additional chunk for the end frames; SAM 3D Body supplies the initial trajectory, pose latents, and initialization and anchor targets. Before each fitting stage, on every dataset, each frame's initializer root is mapped in closed form so that, under the shared camera and body, it keeps its 2-D root projection and projected body height. On 3DPW only, as EMDB does not call this repair and RICH inputs lack its IoU, at the start of every active run of a track after its first, each leading frame whose tracking mask has an IoU below 0.8 with the detection mask, if the next five frames reach 0.8, moves its initializer root along its camera ray to the Theil–Sen extrapolation of those five root depths, unless by under 10 cm.

Implementation details. For the AMASS decoder, pose windows have $N = 45$ frames, each encoded into nine 256-dimensional latent tokens, and consecutive windows of a segment share at least $K = 12$ frames; the Full Model's VAE uses 49-frame windows with seven tokens. Solver chunks have $N' = 25$ frames and share $K' = 9$.

Pose decoder adaptation. The pose decoder $\mathcal{D}$ (7.8M parameters) uses the MLD (Chen et al., 2023) motion-VAE architecture, as in DART (Zhao et al., 2025), trained on AMASS (Mahmood et al., 2019) as 76-dimensional MHR body articulation at 30 fps with $L = 9$ latent tokens of dimension 256 per 45-frame window; the Full Model of the main paper's Tables 1 and 2 instead uses a VAE of the same architecture pretrained on a larger private motion-capture corpus, with 49-frame windows and seven latent tokens. Track segments of at least one window are split into windows $\mathcal{W}_r$ with stride 33, the last end-aligned, and $\mathcal{B}$ blends the decoded windows, halving overlap weights,

into $\widehat{\boldsymbol{\theta}}$. Each person has one decoder copy; segments may bridge inactive gaps of up to 12 frames, filled from the SAM 3D Body trajectory $\boldsymbol{\theta}_{\text{init}}$ without loss, and frames outside segments are optimized directly in registration and fixed in refinement. The initialization stage encodes each window of $\boldsymbol{\theta}_{\text{init}}$ once into frozen-encoder posterior means $\mathcal{Z}_r = \{\mathbf{z}_r^\ell\}_{\ell=1}^L$, keeps them fixed, and fits each person's decoder with its own Adam (Kingma & Ba, 2015) optimizer to the person's mean squared radian error to $\boldsymbol{\theta}_{\text{init}}$, summed over segments, until the person's mean absolute degree error meets the threshold or cap of Table 4, restoring that decoder's best iterate. Both stages then optimize latents and decoders, refinement continuing from registration, with no loss toward pretrained weights or prior.

Objective terms. The data objective is $\mathcal{L}_{\text{data}} = \lambda_{\text{skel}}\mathcal{L}_{\text{skel}} + \lambda_{\text{surf}}\mathcal{L}_{\text{surf}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}$. Landmark $j \in \mathcal{J}_q$, $q \in \{\text{skel}, \text{surf}\}$, of person k in frame $t \in \mathcal{T}_k$ has location $\mathbf{u}_{q,k,j}^t$, confidence $\sigma_{q,k,j}^t$, and camera-space MHR locator or surface point $\widehat{\mathbf{b}}_{q,k,j}^t$, over all 595 surface and 64 of the 70 body-and-hand skeleton landmarks, indices 0–62 and 69; with projection Π^t onto the $W^t \times H^t$ frame and $\mathbf{A}^t = \text{diag}(1/W^t, 1/H^t)$,

$$\mathcal{L}_q = \sum\nolimits_k \tfrac{1}{N_{q,k}} \sum\nolimits_{t\in\mathcal{T}_k,\, j\in\mathcal{J}_q} \mathbf{1}[\sigma_{q,k,j}^t > \tau]\, \mu_{q,j}\, \sigma_{q,k,j}^t \left\| \mathbf{A}^t \big(\Pi^t(\widehat{\mathbf{b}}_{q,k,j}^t) - \mathbf{u}_{q,k,j}^t\big) \right\|_1, \tag{3}$$

where $\mu_{q,j}$ is 4 for hand skeleton landmarks and 1 otherwise, $\tau = 0.3$, $N_{q,k}$ counts valid landmarks, clamped to one, and confidences are zeroed before fitting outside the gate, the SAM 3 mask ω_k^t dilated by 5 pixels, minus crop padding and the occluder mask o_k^t of valid pixels outside ω_k^t more than 0.04 m closer than the median valid MoGe-2 depth inside it, given at least 64 such depths. With ν^t the valid-normal mask that Appendix B.2 defines, b_k^t the $1.05\times$ projected ground-truth-mesh box, absent on frames the dataset marks invalid, Appendix C.1, and $\bar{\omega}_k^t = \bigcup_{k'\neq k} \omega_{k'}^t$, the silhouette target is $\mathbf{s}_k^t = b_k^t\, \omega_k^t\, \nu^t\, (1 - \bar{\omega}_k^t)\, (1 - o_k^t)$ inside the image, built per chunk and never stitched. Its support $\Omega_{\text{sil},k}^t$ is b_k^t minus crop padding, other people's silhouettes, and ignored pixels, namely those outside the crop and those with valid normals outside every trusted tracking mask; rows with a nonempty box but an empty class 1 are written all-ignore (class 2) and contribute nothing. The fixed depth support is $\Omega_{\text{depth}}^t = \bigcup_k \mathbf{s}_k^t$, and at every iteration the normal support Ω_{normal}^t drops from it the rendered pixels of those of the 27 body parts that have skeleton landmarks, assigned by locator parent joint, all with gated confidence at most 0.25. DRTK (Pidhorskyi et al., 2024) renders silhouettes $\widehat{\mathbf{s}}_k^t$ and, with one z-buffer, normals $\widehat{\mathbf{n}}^t$ and depth $\widehat{\mathbf{d}}^t$ on the $W_c \times H_c$ grid of chunk c with frames $\mathcal{C}_c$, pixels i, frame weights w_c^t halved per overlap, $N_w = \sum_c \sum_{t\in\mathcal{C}_c} w_c^t$, active people $\mathcal{A}^t$, chunk c's depth $\tilde{\mathbf{d}}_c^t = a_c \mathbf{d}^t + b_c$ in one sequence-wide gauge, and one global depth scale $\widehat{\alpha}$ and offset $\widehat{\delta}$:

$$\mathcal{L}_{\text{sil}} = \tfrac{1}{N_w} \sum\nolimits_{c,\, t\in\mathcal{C}_c} \tfrac{w_c^t}{H_c W_c \max(1, |\mathcal{A}^t|)} \sum\nolimits_{k\in\mathcal{A}^t,\, i} \Omega_{\text{sil},k}^t(i)\, \big|\widehat{\mathbf{s}}_k^t(i) - \mathbf{s}_k^t(i)\big|, \tag{4}$$

$$\mathcal{L}_{\text{normal}} = \tfrac{1}{N_w} \sum\nolimits_{c,\, t\in\mathcal{C}_c} \tfrac{w_c^t}{3 H_c W_c} \sum\nolimits_i \big\|\Omega_{\text{normal}}^t(i)\big(\widehat{\mathbf{n}}^t(i) - \mathbf{n}^t(i)\big)\big\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{depth}} = \tfrac{1}{N_{\text{depth}}} \sum\nolimits_{c,\, t\in\mathcal{C}_c,\, i} w_c^t\, \Omega_{\text{depth}}^t(i)\, \big|\widehat{\mathbf{d}}^t(i) - (\widehat{\alpha}\, \tilde{\mathbf{d}}_c^t(i) + \widehat{\delta})\big|, \tag{6}$$

with predicted maps $\mathbf{n}^t$, not renormalized, and $\mathbf{d}^t$, and support area N_{depth}. After registration, one robust least-squares solve (Ranftl et al., 2022) on the visible body, the eroded silhouette union with valid normals outside occluders, fixes the chunk affines (a_c, b_c) and initializes $\widehat{\alpha}$ and $\widehat{\delta}$ from chunk-to-render fits and symmetric fits between adjacent chunks on shared pixels; refinement optimizes only $\widehat{\alpha}$ and $\widehat{\delta}$.

Regularization. The regularizer is $\mathcal{L}_{\text{reg}} = \lambda_{\text{zero}}\mathcal{L}_{\text{zero}} + \lambda_{\text{lim}}\mathcal{L}_{\text{lim}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}$, with

$$\mathcal{L}_{\text{zero}} = \sum\nolimits_k \tfrac{1}{76|\mathcal{T}_k|} \sum\nolimits_{t\in\mathcal{T}_k} \big\|\widehat{\boldsymbol{\theta}}_k^t\big\|_2^2, \tag{7}$$

$$\mathcal{L}_{\text{anchor}} = \sum\nolimits_k \tfrac{1}{76|\mathcal{U}_k|} \sum\nolimits_{t\in\mathcal{U}_k} \big\|\widehat{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{\text{init},k}^t\big\|_2^2, \tag{8}$$

$$\mathcal{L}_{\text{lim}} = \sum\nolimits_k \tfrac{1}{|\mathcal{T}_k|} \sum\nolimits_{t\in\mathcal{T}_k} \sum\nolimits_{i=1}^{198} \eta_i\big([\chi_{i,\min} - \widehat{\chi}_{k,i}^t]_+^2 + [\widehat{\chi}_{k,i}^t - \chi_{i,\max}]_+^2\big), \tag{9}$$

$$\mathcal{L}_{\text{inv}} = \sum\nolimits_k \tfrac{1}{N_{\text{inv},k}} \sum\nolimits_{t\in\mathcal{T}_k,\, j} \rho_{k,j}^t\, \pi_j\, \mathbf{1}[\sigma_{\text{skel},k,j}^t < 0.5]\, \big\|\widehat{\mathbf{l}}_{k,j}^t - \mathbf{l}_{\text{init},k,j}^t\big\|_2^2. \tag{10}$$

The trajectory anchor $\mathcal{L}_{\text{anchor}}$ uses anchor frames $\mathcal{U}_k \subseteq \mathcal{T}_k$ that exclude lost-tracking frames, so a person without them contributes zero; refinement anchors to $\widehat{\boldsymbol{\theta}}_{\text{coarse}}$ instead of $\boldsymbol{\theta}_{\text{init}}$. The parameter limit $\mathcal{L}_{\text{lim}}$ covers the 130 local-pose and 68 bone-length coefficients $\widehat{\boldsymbol{\chi}}_k^t$, with MHR bounds and weights η_i. The invisible-landmark anchor $\mathcal{L}_{\text{inv}}$ acts on skeleton landmarks $j \in \{0,\dots,69\}$: $\widehat{\mathbf{l}}_{k,j}^t$ is the canonical locator position and $\mathbf{l}_{\text{init},k,j}^t$ the stop-gradient position under the initializer pose with the current identity and bone lengths; $\rho_{k,j}^t$ marks a reliable initializer body or hand group and is zero on lost-tracking frames, π_j equalizes body-part weights, and $N_{\text{inv},k}$ counts reliable initializer frames, clamped to one.

The smoothness term $\mathcal{L}_{\text{smooth}}$ differs between the stages. Registration uses velocities and accelerations of the root translation $\widehat{\mathbf{t}}^t$ and rotation $\widehat{\mathbf{R}}^t$ and of the local translation $\widehat{\mathbf{a}}_j^t$ and rotation $\widehat{\mathbf{Q}}_j^t$ of joint j from forward kinematics. With $d(\mathbf{R}_1,\mathbf{R}_2) = \|\log(\mathbf{R}_1^\top\mathbf{R}_2)\|_2$ and the SLERP midpoint $\text{mid}(\mathbf{R}_1,\mathbf{R}_2) = \mathbf{R}_1\exp\big(\tfrac{1}{2}\log(\mathbf{R}_1^\top\mathbf{R}_2)\big)$, its terms are

$$\mathcal{L}_{\text{trans}}^{\text{vel}} = \tfrac{1}{3}\,\text{mean}_t\,\big\|\widehat{\mathbf{t}}^{t+1} - \widehat{\mathbf{t}}^t\big\|_2^2, \qquad \mathcal{L}_{\text{root}}^{\text{vel}} = \text{mean}_t\, d\big(\widehat{\mathbf{R}}^t, \widehat{\mathbf{R}}^{t+1}\big)^2, \tag{11}$$

$$\mathcal{L}_{\text{local}}^{\text{vel}} = \text{mean}_{t,j}\,\big[\|\widehat{\mathbf{a}}_j^{t+1} - \widehat{\mathbf{a}}_j^t\|_2^2 + d\big(\widehat{\mathbf{Q}}_j^t, \widehat{\mathbf{Q}}_j^{t+1}\big)^2\big], \tag{12}$$

$$\mathcal{L}_{\text{root}}^{\text{acc}} = \text{mean}_t\, d\big(\widehat{\mathbf{R}}^t, \text{mid}(\widehat{\mathbf{R}}^{t-1}, \widehat{\mathbf{R}}^{t+1})\big)^2, \tag{13}$$

$$\mathcal{L}_{\text{local}}^{\text{acc}} = \text{mean}_{t,j}\,\big[\|\widehat{\mathbf{a}}_j^{t+1} - 2\widehat{\mathbf{a}}_j^t + \widehat{\mathbf{a}}_j^{t-1}\|_2^2 + d\big(\widehat{\mathbf{Q}}_j^t, \text{mid}(\widehat{\mathbf{Q}}_j^{t-1}, \widehat{\mathbf{Q}}_j^{t+1})\big)^2\big], \tag{14}$$

where each mean runs over one contiguous active run, and each term is summed over runs and people. Refinement uses only residuals of a fixed 7-frame, order-two Savitzky–Golay (SG) filter (Savitzky & Golay, 1964) $(h_{-3},\dots,h_3) = \frac{1}{21}(-2,3,6,7,6,3,-2)$, through which gradients flow. On a set of 3D trajectories $\mathcal{Y} = \{\mathbf{y}_i^t\}$,

$$\mathcal{R}_{\text{SG}}(\mathcal{Y}) = \tfrac{1}{N_{\mathcal{Y}}} \sum\nolimits_{t,i} \Big\|\mathbf{y}_i^t - \sum\nolimits_{\Delta=-3}^{3} h_\Delta \mathbf{y}_i^{t+\Delta}\Big\|_2^2, \tag{15}$$

$$\begin{aligned}\mathcal{L}_{\text{smooth}} = {} & \lambda_{\text{SG},1}\mathcal{R}_{\text{SG}}(\{\widehat{\mathbf{m}}_j^t\}) + \lambda_{\text{SG},2}\mathcal{R}_{\text{SG}}(\{\widehat{\mathbf{R}}^t\widehat{\mathbf{m}}_j^t + \widehat{\mathbf{t}}^t\}) \\ & + \lambda_{\text{SG},3}\mathcal{R}_{\text{SG}}(\{\widehat{\mathbf{q}}_{j,e}^t\}) + \lambda_{\text{SG},4}\mathcal{R}_{\text{SG}}(\{\widehat{\mathbf{R}}^t\widehat{\mathbf{q}}_{j,e}^t\}),\end{aligned} \tag{16}$$

over all joints equally, with canonical joint positions $\widehat{\mathbf{m}}_j^t$, rows $\widehat{\mathbf{q}}_{j,e}^t$ of $\widehat{\mathbf{Q}}_j^t$, and the $N_{\mathcal{Y}}$ valid interior positions; shorter sequences use the largest valid odd window or none.

Optimization schedule.

Table 4: Learning rates (LR) and loss weights: a dash marks an inactive term, "fixed" a non-optimized variable, an arrow a scheduled change; initialization uses the Registration column.

Setting	Registration	Refinement
Iterations	600	500
LR root rotation / root translation	0.09 / 1.0	0.09 / 0.04
LR direct body / hand articulation	0.014 / 0.014	fixed / 3×10^{-3}
LR identity / bone-length coefficients	10^{-3} / 10^{-3}	10^{-3} / 10^{-3}
LR focal length	fixed	1.0
LR depth scale $\widehat{\alpha}$ / offset $\widehat{\delta}$ of Appendix A.1	–	5×10^{-3}
LR pose decoder $\mathcal{D}$	10^{-4}	10^{-4}
LR pose latents	10^{-3}	10^{-3}
Initialization stage: LR decoder / latents	10^{-3} / fixed	–
Initialization stage: iteration cap / stop threshold	5,000 / 0.25°	–
Skeleton landmarks (hand multiplier)	1.0 ($\times 4$)	1.0 ($\times 4$)
Surface landmarks	4.0	4.0
Landmark confidence threshold	0.3	0.3
Landmark multiplier, first 25%	$14 \rightarrow 1$	$14 \rightarrow 1$
Normals / silhouette	– / –	5.0 / 1.0
Normal multiplier, first 25%	–	$1/3 \rightarrow 1$
Affine-aligned depth	–	4×10^{-4}
Render scale	–	0.5
Motion-prior loss	–	–
Zero pose	10^{-3}	3×10^{-4}
Parameter limit	10^{-4}	10^{-4}
Trajectory anchor (target)	10^{-2} ($\boldsymbol{\theta}_{\text{init}}$)	1.0 ($\widehat{\boldsymbol{\theta}}_{\text{coarse}}$)
Invisible-landmark anchor	$0.05 \rightarrow 0$ over 150 iterations	–
First-order translation / root rotation / local pose	10^{-3} / 0.5 / 0.5	– / – / –
Translation / root-rotation / local-pose acceleration	– / 0.1 / 0.2	– / – / –
SG local position / rotation (7-frame, order 2)	– / –	0.02 / 50
SG global position / rotation	– / –	10^{-3} / 2

EMDB, RICH, and 3DPW share one configuration except the benchmark-specific settings of Appendix C.1. Both stages use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.85$, $\beta_2 = 0.999$, no weight decay, gradient-norm clipping at 5, a 250-iteration linear warm-up from 0.1 times the nominal rate, and cosine decay to 10^{-4}; they skip non-finite steps, and from iteration 10 a loss above twice the last accepted one restores that iterate with its Adam moments and halves a persistent rate multiplier.

The following shared settings were chosen with evaluation data: the reuse of the registered latents in refinement and the initialization-stage stopping rule were adopted after comparisons on the RICH test population, the pose-latent learning rate was chosen from 10^{-2} and 10^{-3} by the lower final refinement objective, without ground truth, on one RICH test sequence, and the render scale and the predictor's number of denoising steps follow the EMDB1+2 ablations of Appendix E. The remaining decoder and latent learning rates were set by hand.

B Predictor Details

B.1 Sapiens2 Body

Sapiens2 Body is PixelBody's landmark predictor; Appendix E.4 evaluates its landmarks as fitting targets, and Appendix E.3 compares later checkpoints from separate training runs with public predictors. It uses a pretrained 0.8B Sapiens2 vision-transformer (ViT) backbone (Khirodkar et al., 2026) with 512×512 input, a heatmap-based landmark prediction head with a separate branch for uncertainty, a segmentation head with PointRend refinement and a feed-forward pose head. The model is jointly trained on all objectives, but the pose head is omitted during inference. It predicts 939 landmarks (70 body-and-hand, 274 face, and 595 surface) and 34 segmentation classes. PixelBody fits all 595 MHR surface landmarks and 64 of the 70 body-and-hand landmarks (22 body, 42 hand), with the hand landmarks weighted $4\times$. Table 5 lists the training stages. No training data comes from EMDB, RICH, or 3DPW footage. Objectives include a smooth-L1 loss on heatmaps, a Gaussian negative log-likelihood (GNLL) loss for uncertainty, focal loss for segmentation, a separate smooth-L1 point loss for PointRend, and L1 losses for pose parameters. We use AdamW (weight

decay 10^{-2}, base learning rate 3×10^{-4}, layer-wise decay 0.85 per ViT block, $3\times$ learning rate and $0.1\times$ weight decay for the heads), fixed learning rate for 70% of the iterations with a 1k-iteration warmup, followed by a cosine decay schedule.

Table 5: Training stages of the Sapiens2 Body landmark and segmentation model; budgets are optimizer steps.

Stage	Content	Budget
Pretraining	Landmarks only; ViT from pretrained 0.8B weights	100k
All tasks	939 landmarks, 33 classes segmentation with point-based mask refinement, pose	60k
Data expansion	Adds datasets with partial ground-truth; trains heads only with frozen visual encoder	60k
Final fine-tuning	unfreezes all parameters	30k

A separately fine-tuned dense checkpoint, the Sapiens2 Body row of Table 13, shares the backbone but uses a multi-scale upsampler over four ViT blocks and predicts, from a single-person crop, 17 segmentation classes, 2 UV channels, metric body depth as a per-person average plus a zero-mean residual, and 3 body-normal channels. It is trained on synthetic MHR renders for 30k frozen-ViT training and 200k full fine-tuning with segmentation, UV, depth, and normal losses.

B.2 WAN2.1 BODY

Wan2.1 Body uses separate Wan 2.1-1.3B (Wan Team, 2025) video diffusion transformers $\mathbf{v}_\psi$ for body depth and normals, each initialized from the text-to-video model with a fixed empty prompt. As in Wan image-to-video, noisy target latents are concatenated along channels with the clean latents of the RGB clip $\mathbf{x}_p$ of predictor window p, and $\mathbf{v}_\psi$ predicts the flow-matching velocity (Lipman et al., 2023; Liu et al., 2023): with $\mathbf{g}_\kappa = (1-\kappa)\mathbf{g}_0 + \kappa\boldsymbol{\epsilon}$ and $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{g}_0$,

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\kappa,\boldsymbol{\epsilon}}\Big[w(\kappa)\,\big\|\mathbf{v}_\psi(\mathbf{g}_\kappa, \mathbf{x}_p, \kappa) - \mathbf{v}\big\|_2^2\Big], \tag{17}$$

averaged over all target-latent positions and channels, where $w(\kappa) \propto \exp\big(-2(\kappa - \frac{1}{2})^2\big) - e^{-1/2}$ is normalized to mean one over the 1,000 schedule levels. We fully fine-tune on BEDLAM (Black et al., 2023) and BEDLAM2 (Tesch et al., 2025) with noise-schedule shift 5 and 25-frame zero-identity targets rendered from MHR fits of their SMPL-X bodies, on one- and multi-person crops including half-body and part views, with three random color, blur, or compression transforms per clip and, for half of the clips, a background patch over part of the body. Both predictors use the Aurora optimizer (Dewulf et al., 2026), a Muon (Jordan et al., 2024) variant (learning rate 8×10^{-4}, Nesterov momentum 0.95, three polar iterations, no weight decay), with an AdamW group (learning rate 2×10^{-5}) for the non-2D parameters, a 1,000-step warmup from $0.1\times$, 20,000 steps on 1280×720 and 720×1280 clips at global batch size 64, gradient clipping at 1, and EMA weights with decay $\min\big(0.9995, (1+t)/(10+t)\big)$ after t updates.

Inference. PixelBody runs the predictors at 720p with two unguided denoising steps from seeded Gaussian noise at shift 5. Each solver chunk is one 25-frame predictor window on all three datasets. The outputs are body-depth maps $\mathbf{d}_p^t$ and normal maps $\mathbf{n}_p^t$ in the right-handed OpenGL camera ($+x$ right, $+y$ up, $+z$ toward the camera), without validity masks. The targets use the zero-identity shape, with identity and face blendshapes zeroed and bone lengths kept, so PixelSolver renders the zero-identity mesh for the normal, depth, and silhouette losses. PixelSolver marks a predicted-normal pixel as valid when its decoded magnitude lies in $[0.6, 1.4]$, after pixels whose 8-bit channels all exceed 245 are reset to mid-gray, so the near-white background is rejected.

C EVALUATION PROTOCOL

C.1 BENCHMARK INPUTS

EMDB1+2 denotes the 39 sequences of the official EMDB1 and EMDB2 evaluation subsets, which contain 17 and 25 sequences and share three. PixelBody uses boxes, crops, or track prompts derived

from ground-truth annotations on all three benchmarks. On EMDB (Kaufmann et al., 2023), boxes around SAM 3 (Carion et al., 2026) tracking masks obtained without ground-truth prompts prompt the SAM 3D Body initializer (Yang et al., 2026b) and the landmark predictor, whereas the solver's tracking masks come from a second SAM 3 track prompted with the projected ground-truth-mesh box at an anchor frame; the dense-predictor and solver render crops are boxes around the projected ground-truth meshes. On RICH (Huang et al., 2022), $1.05\times$ boxes around each subject's visible projected ground-truth mesh prompt the initializer and the landmark predictor in every frame and give the dense-predictor crops; SAM 3 is prompted once per subject. On 3DPW (von Marcard et al., 2018), SAM 3 tracks each person from one $1.05\times$ box around its ground-truth 2D landmarks, and $1.05\times$ boxes over the tracks intersected with per-frame SAM 3 detections prompt the mask-conditioned initializer. Fitting uses RICH frames whose SAM 3 track box overlaps the projected ground-truth box with an IoU of at least 0.05 and 3DPW frames whose track mask overlaps a per-frame SAM 3 detection with a mask IoU of at least 0.05. Sapiens2 Body predicts all fitted landmarks, and ground-truth 2D landmarks are not used. Silhouettes, occluder masks, and landmark gates are built in the same way on all three benchmarks, as Appendix A.1 defines.Frames that the dataset marks invalid, through EMDB's good-frame mask or 3DPW's camera-pose validity flag, get no box and therefore contribute no silhouette, normal, or depth loss of their own, whereas their landmark losses still apply; on frames mixing valid and invalid people, the invalid person's surface still renders into the shared scene render and can receive normal and depth gradients through other people's support. 1.0% of scored EMDB frames and 32.0% of scored 3DPW frames lie within the motion prior's 45-frame window of such a frame. Beyond these inputs, which EMDB loads through its own input path with an end-aligned final chunk, PixelSolver's losses and optimization settings differ across benchmarks only in the re-acquired-track repair, which only 3DPW uses, Appendix A.1.

C.2 Metrics

EMDB and RICH use the standard 24-joint protocol and 3DPW the standard 14-joint protocol; MPJPE, PA-MPJPE, and PVE are in millimeters. World motion on EMDB2 and RICH uses the 100-frame segments of WHAM (Shin et al., 2024): WA-MPJPE$_{100}$ aligns each segment with one similarity transform fitted to all of its frames, and W-MPJPE$_{100}$ with one fitted to its first two frames. The root translation error (RTE) rigidly aligns each whole predicted root trajectory to the ground truth without scaling and averages, over all frames, each frame's root-position error divided by the sequence's total ground-truth root path length, in percent. Jerk, in m/s^3, is the norm of the third-order finite difference of camera-space SMPL (Loper et al., 2015) joint positions times 30^3, over 15 joints without the pelvis, ankles, feet, wrists, and hands, within contiguous runs of good frames, averaged per sequence and then equally over the 39 EMDB1+2 sequences. Tables give Jerk in 10^2 m/s^3, the text in m/s^3; ground-truth motion scores 3.0. The AMASS-decoder model, whose geometry errors appear in Table 7, scores 3.6, and the mask-box SAM 3D Body initializer of that table scores 10.5. Tables rank Jerk by its distance to the ground truth's, not by its value, because over-smoothed motion also scores low.

C.3 Evaluators and Populations

On EMDB, we apply the good-frame mask, align MPJPE and PVE at the hip center (the mean of joints 1 and 2), use a per-frame Umeyama similarity (Umeyama, 1991) for PA-MPJPE, and pool WA-MPJPE$_{100}$ and W-MPJPE$_{100}$ over all 43,129 good EMDB2 frames, as WHAM's (Shin et al., 2024) released code does. On RICH, we use GVHMR's (Shen et al., 2024) population of 191 sequence–camera items and 106,446 frames, with the released per-frame SMPL-X meshes transferred to SMPL as ground truth. The reproduced SAM 3D Body RICH row uses 21,287 person-frames sampled every fifth frame from the same 191 items, so it is not directly comparable with PixelBody's RICH rows. On 3DPW, we score 35,310 of the 35,515 official 14-joint metric-valid person-frames, the other 205 lacking SAM 3 support, whereas GVHMR, GENMO (Li et al., 2025), and the reproduced SAM 3D Body row use the 34,561 person-frames of the WHAM/TCMR (Shin et al., 2024; Choi et al., 2021) 2D-landmark validity mask. The official 3DPW sequence files list all 37 test actors as male, so our 3DPW ground truth uses the male SMPL body, and our 3DPW rows differ from the published rows in the evaluated population described above; the comparison between PixelBody and its initializer uses the same ground truth.

The main paper's Table 1 and Table 6 average over all evaluated frames, or person-frames on 3DPW, as the published baselines do; the controlled EMDB1+2 comparisons, including Table 7, average within each sequence and weight sequences equally. MHR outputs are converted to SMPL by optimization without ground truth, with one neutral-SMPL identity per tracked person; the converter's default pre-pass, used for every MHR-based row we computed, has a frame-indexing error that changes EMDB PA-MPJPE by 0.2–0.4 mm.

C.4 Comparison Sources and SAM 3D Body Initializer

A dash (–) marks an unreported benchmark; the main paper's Tables 1 and 2 rank all entries, including baselines trained on a benchmark, except the gray cells, which are NLF-L+fit on RICH, its training benchmark, and THFM, which predicts joints only, and the RTE column.

In Table 8, a dagger marks EMDB2 values with ground-truth intrinsics, as the cited papers or released code state or imply: ground-truth-camera TRAM and MetricHMSR, and estimated-camera TRAM, GENMO, PromptHMR-vid, DuoMo, MetricHMSR, and HTD-Refine.

PixelBody's predictors and the pretrained VAE were not trained on any of the benchmarks; the decoder is adapted to each test video at inference. The reproduced SAM 3D Body rows use the public DINOv3 (Siméoni et al., 2025) checkpoint with per-sequence mean MoGe-2 intrinsics instead of the initializer's per-frame ones, one shape and scale per sequence and mask-conditioned SAM 3 mask boxes on EMDB, and unconditioned clipped $1.05\times$ ground-truth boxes on RICH and 3DPW, Appendix E.2.

D Full Benchmark Results

D.1 Camera Space

Table 6 extends the main paper's Table 1 with all methods and 3DPW.

On EMDB1, the AMASS-decoder model and the Full Model have the lowest PA-MPJPE, MPJPE, and PVE among the compared published mesh-recovery rows under their reported protocols; the joints-only THFM row has lower MPJPE. With Sapiens2 Body, PixelBody has lower PA-MPJPE and MPJPE than SAM 3D Body-DINOv3 but 0.8 mm higher PVE. The PVE margins over SAM 3D Body-DINOv3 (72.5 mm) are 1.7 mm for the AMASS-decoder model, within the 1.8 mm by which our optimization-based SMPL conversion lowered EMDB1 PVE relative to a linear surface mapping in an earlier comparison, and 2.0 mm for the Full Model, whereas SAM 3D Body maps the MHR mesh to the SMPL mesh format; the PA-MPJPE and MPJPE margins exceed the 0.9–1.2 mm conversion gaps, and the reproduced SAM 3D Body initialization is weaker on all three metrics.

On RICH, NLF-L+fit (trained on RICH) has lower camera-space PA-MPJPE than PixelBody in all three configurations, SAM 3D Body-DINOv3 and SAM 3D Body-H also than the two with the AMASS decoder, and NLF-L+fit, SAM 3D Body-DINOv3, and DuoMo have lower MPJPE and PVE than all three; this ranking is descriptive, as populations and inputs are not matched, and per-frame camera-space metrics do not reflect temporal stability, Appendix C.2.

On 3DPW, PixelBody is worse in every configuration on all three metrics than its initializer setting of Appendix E.2 on the same 35,310 samples and the same ground truth. PixelBody also has higher MPJPE and PVE than most image-based and some video-based methods; these rows differ from the published rows in the evaluated population, and our 3DPW ground truth uses the male SMPL body, as the official 3DPW sequence files list all 37 test actors as male, Appendix C.3. SAM 3D Body's training data include 3DPW, and our SAM 3D Body reproduction and PixelBody use different 3DPW populations, preprocessing, and pipelines, so their 3DPW comparison is descriptive rather than a controlled ablation.

D.2 EMDB1+2

Table 7 adds two SAM 3D Body reproductions on EMDB1+2: the mask-box initializer of the main paper's Table 3a, under the protocol of Appendix C like the AMASS-decoder model, and the body-

Table 6: Camera-space errors (mm, frame-averaged) on EMDB1 and 3DPW (top) and RICH (bottom), sorted by year within each group. PA: PA-MPJPE; †: ground-truth intrinsics, for all values when on a name; our run: released code run by us; ‡: 3D joints only, no mesh; –: not reported.

	EMDB1 (24)			3DPW (14)		
Method	**PA↓**	**MPJPE↓**	**PVE↓**	**PA↓**	**MPJPE↓**	**PVE↓**
Image-based methods						
HMR2.0b (Goel et al., 2023)	79.2	118.5	140.6	54.3	81.3	93.1
NLF-L+fit (Sárándi & Pons-Moll, 2024)	40.9	68.4	80.6	33.6	54.9	63.7
CameraHMR (Patel & Black, 2025)	43.3	70.3	81.7	35.1	56.0	65.9
PromptHMR† (Wang et al., 2025b)	41.0	71.7	84.5	36.6	58.7	69.4
SAM 3D Body-H (Yang et al., 2026b)	38.5	62.9	74.3	**33.2**	54.8	64.1
SAM 3D Body-DINOv3 (Yang et al., 2026b)	38.2	61.7	72.5	33.8	54.8	63.6
SAM 3D Body, our run (Yang et al., 2026b)	39.9	63.0	75.9	33.8	55.8	66.5
MetricHMSR (Song et al., 2026)	43.2	70.7	81.6	33.6	**53.0**	**62.7**
Multi-HMR 2 (Fiche et al., 2026)	48.3	68.5	78.7	40.5	68.2	80.1
Video-based methods						
GVHMR (Shen et al., 2024)	42.7	72.6	84.2	36.2†	55.6†	67.2†
TRAM (Wang et al., 2024)	45.7†	74.4†	86.6†	35.6	59.3	69.6
WHAM (Shin et al., 2024)	50.4	79.7	94.4	35.9	57.8	68.7
GENMO (Li et al., 2025)	42.5	73.0	84.8	34.6†	53.9†	65.8†
DuoMo† (Wang et al., 2026d)	41.8	67.1	78.2	–	–	–
Human3R (Chen et al., 2026)	48.5	73.9	86.0	44.1	71.2	84.9
OnlineHMR (Zhao et al., 2026)	46.0	74.0	86.1	43.7	69.9	83.7
THFM‡ (Wang et al., 2026a)	35.2	54.6	–	–	–	–
SHOW (Shi et al., 2026)	45.8	75.7	88.8	41.0	67.7	78.7
UniSH (Li et al., 2026a)	86.6	112.9	140.0	48.8	75.6	88.8
Ours (Sapiens2 Body, AMASS decoder)	35.8	61.3	73.3	36.0	59.7	70.7
Ours (Wan2.1 Body, AMASS decoder)	**35.1**	59.1	70.8	35.7	58.6	69.9
Ours (Full Model)	35.2	**59.0**	**70.5**	35.1	57.6	68.5

	RICH (24)		
Method	**PA↓**	**MPJPE↓**	**PVE↓**
Image-based methods			
HMR2.0b	48.1	96.0	110.9
NLF-L+fit	28.7	51.0	58.2
CameraHMR	34.0	55.7	64.4
PromptHMR†	37.3	56.6	65.5
SAM 3D Body-H	31.9	55.0	61.7
SAM 3D Body-DINOv3	30.9	53.7	60.3
SAM 3D Body, our run	32.5	55.8	63.2
Video-based methods			
GVHMR	39.5†	66.0†	74.4†
WHAM	44.3	80.0	91.2
GENMO	39.1†	66.8†	75.4†
DuoMo†	34.8	**51.4**	**58.9**
THFM‡	34.7	55.0	–
Ours (Sapiens2 Body, AMASS decoder)	33.4	59.4	67.1
Ours (Wan2.1 Body, AMASS decoder)	32.2	57.4	65.0
Ours (Full Model)	**30.6**	54.8	61.7

mode, mask-conditioned row of Table 10 with ground-truth-mesh boxes, like PixelBody's crops, scored on the same good frames without sequence-shared identity and bone length.

Table 7: Camera-space errors (mm) on EMDB1+2 (24 joints). Our run: the public DINOv3 checkpoint run by us with optimization-based SMPL conversion; ground-truth-mesh boxes: Table 10.

Method	**PA-MPJPE ↓**	**MPJPE ↓**	**PVE ↓**
SAM 3D Body, our run (Yang et al., 2026b)	36.1	56.5	69.0
SAM 3D Body, our run (Yang et al., 2026b), ground-truth-mesh boxes	33.3	53.4	64.3
Ours (Sapiens2 Body, AMASS decoder)	32.3	53.4	65.3
Ours (Wan2.1 Body, AMASS decoder)	31.9	52.4	64.1
Ours (Full Model)	**31.8**	**51.9**	**63.2**

D.3 World Space

Table 8 compares all methods in world space; Appendix C.2 defines the metrics. PixelBody evaluates consecutive source frames, pools frames across items, and uses clipped $1.05\times$ boxes projected from ground-truth meshes; the published methods share no common protocol. Rows are grouped into feed-forward, causal online, and offline methods and sorted by year; ground-truth-camera rows form a separate protocol. Estimated-camera rows use each method's reported cameras; ours use DROID-W (Li et al., 2026b) or Depth Anything 3 (DA3; Lin et al., 2026) on EMDB2 and optimization-based MHR-to-SMPL conversion, and all rows we computed are frame-pooled. The released JOSH code omits the paper's 2D-keypoint term, so our JOSH run mainly reflects its VIMO initialization. HTD-Refine refines GVHMR outputs and reports RTE in metres, so its RTE cells are empty; RICH rows omit camera-source suffixes because the fixed cameras share a global frame.

E Additional Ablations

E.1 Solver Ablations

The main paper's Table 3 gives world-space errors for every ablation row; with ground-truth cameras, adding depth to normals lowers W-MPJPE$_{100}$ from 126.0 to 124.6 mm.

With 95% CIs from 100,000 bootstrap samples clustered by subject, the PA-MPJPE reductions of decoder adaptation, $[-2.83, -1.02]$ mm in registration and $[-2.33, -0.80]$ mm in refinement, and of fitting from the GT-box initializer, $[-1.92, -0.70]$ mm, exclude zero, whereas every MPJPE and PVE interval of these pairs, and every interval of adapting both blocks instead of the decoder alone, includes zero. Every row changes only the ablated setting of the AMASS-decoder model: in panel (b), the adapted-decoder rows are its two stages and the direct-pose rows one run with direct pose optimization in both, and in panel (c), adapting both blocks is the AMASS-decoder model. The mask-box initializer row is our SAM 3D Body (Yang et al., 2026b) reproduction before fitting, as Appendix C.4 describes, with identity and bone length shared across each sequence; without this sharing, it has a PA-MPJPE, MPJPE, and PVE of 36.0, 55.9, and 68.3 mm and a Jerk of 10.4. Jerk is 10.5 with this sharing, 4.7 and 5.6 for direct-pose and adapted-decoder registration, and 3.0 and 3.6 for the matched direct-pose and adapted-decoder refinement rows. The GT-box initializer row is the body-mode, mask-conditioned row of Table 10, with per-frame identity and bone length. The latent-only setting freezes the pose decoder in every stage, so its initialization stage fits the latents.

Table 9 re-runs the AMASS-decoder model's refinement from the same registration at full and half render resolution, timed on one H200 GPU as wall time including input loading, whereas Appendix F times stages from their completion markers. At half resolution, each geometry error changes by at most 0.18 mm and Jerk by -1.63 m/s^3, and refinement runs $1.50\times$ faster; an unchanged re-run of the AMASS-decoder model's refinement moves the means by up to 0.06 mm and single sequences by up to 3.4 mm.

Table 8: World-space errors on EMDB2 and RICH (mm; RTE in % of root path length), 100-frame windows, 24 joints. †: ground-truth intrinsics on EMDB2; *: external camera pipeline on EMDB2; our run: released code run by us; §: PixelSolver's starting point, lifted with the same cameras.

	EMDB2 (24)			RICH (24)		
Method	**WA-MPJPE$_{100}$↓**	**W-MPJPE$_{100}$↓**	**RTE↓**	**WA-MPJPE$_{100}$↓**	**W-MPJPE$_{100}$↓**	**RTE↓**
GT cameras						
TRAM† (Wang et al., 2024)	62.7	182.8	0.2	–	–	–
GENMO (Li et al., 2025)	69.5	185.9	0.9	–	–	–
PromptHMR-vid (Wang et al., 2025b)	52.9	142.1	0.2	–	–	–
DuoMo, our run (Wang et al., 2026d)	52.8	136.1	0.2	–	–	–
JOSH, our run (Liu et al., 2026)	67.1	183.4	0.2	–	–	–
MetricHMSR† (Song et al., 2026)	55.6	152.5	0.2	–	–	–
SAM 3D Body, our run	51.5	134.7	0.2	–	–	–
PixelBody start§	50.8	134.0	0.2	–	–	–
Ours (Sapiens2 Body, AMASS decoder)						
GT cameras	46.1	125.4	0.2	–	–	–
Ours (Wan2.1 Body, AMASS decoder)						
GT cameras	45.3	124.6	0.2	–	–	–
Ours (Full Model)						
GT cameras	**44.7**	**122.6**	0.2	–	–	–
Feed-forward						
GENMO*† (Li et al., 2025)	74.3	202.1	1.2	75.3	118.6	1.9
MetricHMSR† (Song et al., 2026)	72.1	199.5	1.4	109.6	165.5	1.9
SHOW (Shi et al., 2026)	109.1	262.3	2.1	107.3	172.7	2.2
UniSH (Li et al., 2026a)	118.5	270.1	5.8	118.1	183.2	4.8
Online (causal)						
WHAM (Shin et al., 2024)	135.6	354.8	6.0	109.9	184.6	4.1
Human3R (Chen et al., 2026)	112.2	267.9	2.2	110.0	184.9	3.3
JOSH3R (Liu et al., 2026)	220.0	661.7	13.1	190.4	334.9	6.3
OnlineHMR (Zhao et al., 2026)	93.5	310.4	2.2	–	–	–
Offline						
GVHMR* (Shen et al., 2024)	111.0	276.5	2.0	78.8	126.3	2.4
TRAM† (Wang et al., 2024)	76.4	222.4	1.4	127.8	238.0	6.0
PromptHMR-vid*† (Wang et al., 2025b)	71.0	216.5	1.3	118.2	207.0	5.6
DuoMo*† (Wang et al., 2026d)	**67.6**	**171.6**	1.2	53.6	80.8	1.3
EmbodMocap (Wang et al., 2026c)	82.2	220.7	1.7	–	–	–
HTD-Refine*† (Wei et al., 2026)	69.2	192.4	–	75.2	123.8	–
HumanBA (Yang et al., 2026a)	70.1	196.0	1.2	–	–	–
JOSH (Liu et al., 2026)	68.9	174.7	1.3	102.6	184.3	3.4
Ours (Sapiens2 Body, AMASS decoder)						
DA3	75.1	215.6	1.7	–	–	–
DROID-W	72.7	211.0	1.7	55.5	85.9	1.2
Ours (Wan2.1 Body, AMASS decoder)						
DA3	73.9	213.5	1.7	–	–	–
DROID-W	71.2	207.8	1.7	52.5	81.6	1.2
Ours (Full Model)						
DA3	73.1	212.1	1.7	–	–	–
DROID-W	70.3	205.4	1.7	**49.7**	**77.7**	1.2

Table 9: Refinement at full and half render resolution on EMDB1+2. The half-resolution row reruns the AMASS-decoder model's refinement, and the full-resolution row changes only the render scale.

Resolution	**PA-MPJPE** (mm)	**MPJPE** (mm)	**PVE** (mm)	**Jerk** (10^2 m/s^3)	**Time** (s/frame)
Full	**31.97**	**52.41**	64.27	3.56	1.97
Half	**31.97**	52.42	**64.09**	**3.54**	**1.32**

E.2 INITIALIZER

Table 10 runs SAM 3D Body in full mode, with hand refinement, and body mode, each with and without SAM 3 mask conditioning, on shared frames with per-sequence mean MoGe-2 intrinsics, oracle associations, and $1.05\times$ ground-truth-mesh boxes. Removing the mask changes every metric by at most 0.5 mm. These EMDB rows are not comparable with the main paper's mask-box initializer row, which uses SAM 3 mask boxes obtained without ground-truth prompts and shares identity and bone length across each sequence.

Table 10: SAM 3D Body accuracy in millimeters by mode and SAM 3 mask conditioning, over EMDB's 62,425 good frames, averaged per sequence first, and 3DPW's 35,310 mask-supported samples.

		EMDB (24 joints)			**3DPW (14 joints)**		
Mode	**SAM 3 mask**	**PA-MPJPE ↓**	**MPJPE ↓**	**PVE ↓**	**PA-MPJPE ↓**	**MPJPE ↓**	**PVE ↓**
Full	✓	34.7	54.3	66.1	34.1	55.5	66.2
Full	–	34.7	54.3	66.1	34.1	56.0	66.6
Body	✓	**33.3**	**53.4**	**64.3**	**33.5**	**55.2**	**65.3**
Body	–	33.4	53.5	64.4	33.6	55.7	65.7

E.3 LANDMARK PREDICTOR EVALUATION

Table 11 recomputes every metric from public weights on single-person crops from $1.2\times$ ground-truth boxes: COCO-17 (Lin et al., 2014) skeleton joints on RICH and EMDB against ViTPose++ (Xu et al., 2022; 2024) and Sapiens2 (Khirodkar et al., 2026), and MammaNet's 512 SMPL-X (Pavlakos et al., 2019) surface landmarks on RICH, EMDB, and MammaEval-S (Cuevas-Velasquez et al., 2026), which Sapiens2 Body predicts through a comparison head trained on BEDLAM (Black et al., 2023) over its frozen backbone. PCKh@τ (Andriluka et al., 2014) counts joints within τ head sizes, a head size being 0.6 times the projected head-vertex box diagonal, and MammaNet* is the released MammaNet under this protocol, since MAMMA's evaluator was unavailable to us.

Table 11: Landmark accuracy, with surface errors in crop-space pixels, of public predictors and of later Sapiens2 Body checkpoints, not the one PixelBody fits.

Model	**RICH**	**EMDB**	**MammaEval-S**
Skeleton landmarks (PCKh@0.1 / 0.5 ↑)			
ViTPose++-H (Xu et al., 2024)	39.7% / 96.4%	42.1% / 97.6%	–
Sapiens2-1B-Pose (Khirodkar et al., 2026)	42.9% / 97.4%	46.9% / 98.4%	–
Sapiens2 Body (ours)	**51.0% / 98.9%**	**53.0% / 99.2%**	–
Surface landmarks (pixel error ↓)			
MammaNet* (Cuevas-Velasquez et al., 2026)	8.25	8.96	6.62
Sapiens2 Body (ours)	**7.43**	**7.60**	**6.35**

E.4 LANDMARK SOURCES

Table 12: The AMASS-decoder model's registration stage on EMDB1+2 by landmark source, in millimeters, Jerk (10^2 m/s^3) ranked by distance to the ground truth's 3.0. Row 1 is the main paper's Table 3b adapted-decoder registration row.

Skeleton landmarks	Surface landmarks	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Jerk
Sapiens2 Body	Sapiens2 Body	33.3	55.2	67.5	5.6
	MammaNet	33.9	55.3	67.5	6.0
	None	35.4	58.3	70.4	5.7
Sapiens2 Body-17	Sapiens2 Body	33.1	54.7	66.6	5.1
	MammaNet	33.6	54.9	66.7	5.6
	None	35.3	57.5	68.8	4.6
ViTPose-17	Sapiens2 Body	33.8	55.0	67.1	5.1
	MammaNet	37.0	57.3	70.4	6.1
	None	42.8	63.7	76.7	6.8
None	Sapiens2 Body	**32.4**	**54.4**	**66.5**	4.9
	MammaNet	35.1	56.3	68.8	6.2
	None	65.1	136.5	161.9	**1.9**

Table 12 swaps the landmark sources of the AMASS-decoder model's registration stage on EMDB1+2. Sapiens2 Body supplies 64 of its 70 body-and-hand skeleton landmarks and all 595 surface landmarks, Sapiens2 Body-17 is their COCO-17 subset, ViTPose-17 uses ViTPose-B (Xu et al., 2022), and MammaNet's 512 SMPL-X landmarks map to MHR through a barycentric correspondence; "None" disables a reprojection term but not the invisible-landmark anchor. Every row is the AMASS-decoder model's registration stage with only the landmark sources changed, except that MammaNet cells fit without identity blendshapes, which its template-surface correspondence does not support. Replacing one Sapiens2 Body source at a time, by MammaNet surface landmarks or by Sapiens2 Body-17 or ViTPose-17 skeleton landmarks, changes PA-MPJPE by at most 0.6 mm. Either landmark type can thus be replaced by a public predictor at a PA-MPJPE cost of at most 0.6 mm in the registration stage. In the registration stage, removing the skeleton term while keeping Sapiens2 Body surface landmarks lowers each geometry metric by up to 0.9 mm and lowers Jerk; PixelBody keeps the term because it carries the 42 hand landmarks, which whole-body metrics do not score.

E.5 DENSE PREDICTORS

In Table 13, each Wan2.1 Body row changes both depth and normals, and the 480p and 720p predictors also differ in training schedule. Beyond two function evaluations (NFE), geometry errors change by at most 0.1 mm while throughput falls, consistent with the low-NFE saturation of prior dense-prediction work (Gui et al., 2025; Xu et al., 2025; Garcia et al., 2025; He et al., 2025; Zhang et al., 2026a). A 480p ablation fine-tunes Wan 2.1-1.3B on 25-frame clips for 10,000 steps at batch size 128 and training shift 5 or 12, scored with EMA weights at inference shift 12 on 59 held-out BEDLAM clips. At NFE 2, the normal angular errors are $11.68°$ and $11.69°$ for training shifts 5 and 12; neither another NFE from 3 to 30 nor the other shift lowers them significantly, as every paired 95% CI includes zero and none survives Holm correction, so we keep training shift 5 and NFE 2.

Table 13: EMDB1+2 results by depth and normal predictor, in millimeters; Jerk (10^2 m/s^3) is ranked by distance to the ground truth's 3.0; FPS: normal predictor, one A100. The starred NFE is PixelBody's; its row is not the AMASS-decoder model, and all runs predate its 25-frame chunks.

Backbone	Size	Res.	NFE	FPS ↑	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Jerk
Wan2.1 Body	1.3B	720p	1	3.07	32.0	52.6	63.6	**3.2**
Wan2.1 Body	1.3B	720p	2*	2.80	**31.7**	**52.4**	63.5	**3.2**
Wan2.1 Body	1.3B	720p	5	1.93	**31.7**	**52.4**	**63.4**	**3.2**
Wan2.1 Body	1.3B	720p	10	1.25	**31.7**	**52.4**	**63.4**	3.3
Wan2.1 Body	1.3B	480p	2	7.14	32.2	53.1	64.1	**3.2**
Wan2.1 Body	1.3B	480p	10	3.52	32.2	53.1	64.1	**3.2**
Sapiens2 Body	0.8B	1024^2	–	**33.68**	33.0	55.0	66.3	3.8

F RUNTIME

Sapiens2 Body is timed on one A100-SXM4-80GB GPU at batch size 4, 1024×1024, bfloat16, without the unused pose head, a setup separate from the FPS column of Table 13. All timings are for offline batch processing. Solver timing is weighted over the EMDB evaluation set with 25-frame solver chunks overlapping by 9 frames, from the end of dense-input loading to each stage's saved result, so registration includes the initialization stage. The A100 run has the AMASS-decoder model's chunking, iteration counts, and render scale but earlier settings otherwise, and its start-up and dense-input loading add 0.37 s per frame; on one H200, the AMASS-decoder model takes 0.06 s and 1.20 s per frame for registration and refinement. Wan2.1 Body normal counts every frame of each 25-frame predictor window, about 1.53 times the reciprocal of the FPS in Table 13 at EMDB's 9-frame overlap.

Table 14: Partial per-frame latency in seconds, profiled per stage; bold subtotals sum the rows and are not end-to-end timings. +Wan2.1 Body normal replaces the Sapiens2 Body dense predictor with the two-NFE, 720p Wan2.1 Body normal predictor; neither includes Wan2.1 Body depth or SAM 3.

Stage	**Component**	**Base**	**+Wan2.1 Body normal**
Preprocessing	MoGe-2	0.14	0.14
	Sapiens2 Body (dense)	0.06	–
	Wan2.1 Body normal	–	0.55
	Sapiens2 Body (landmarks)	0.17	0.17
	SAM 3D Body	0.21	0.21
Solver	Registration	0.12	0.12
	Refinement	2.17	2.17
Subtotal		**2.87**	**3.36**

G QUALITATIVE RESULTS

Figure 6 shows the stages of PixelBody's fit on the main paper's Figure 4 clips, and Figure 7 failure cases.

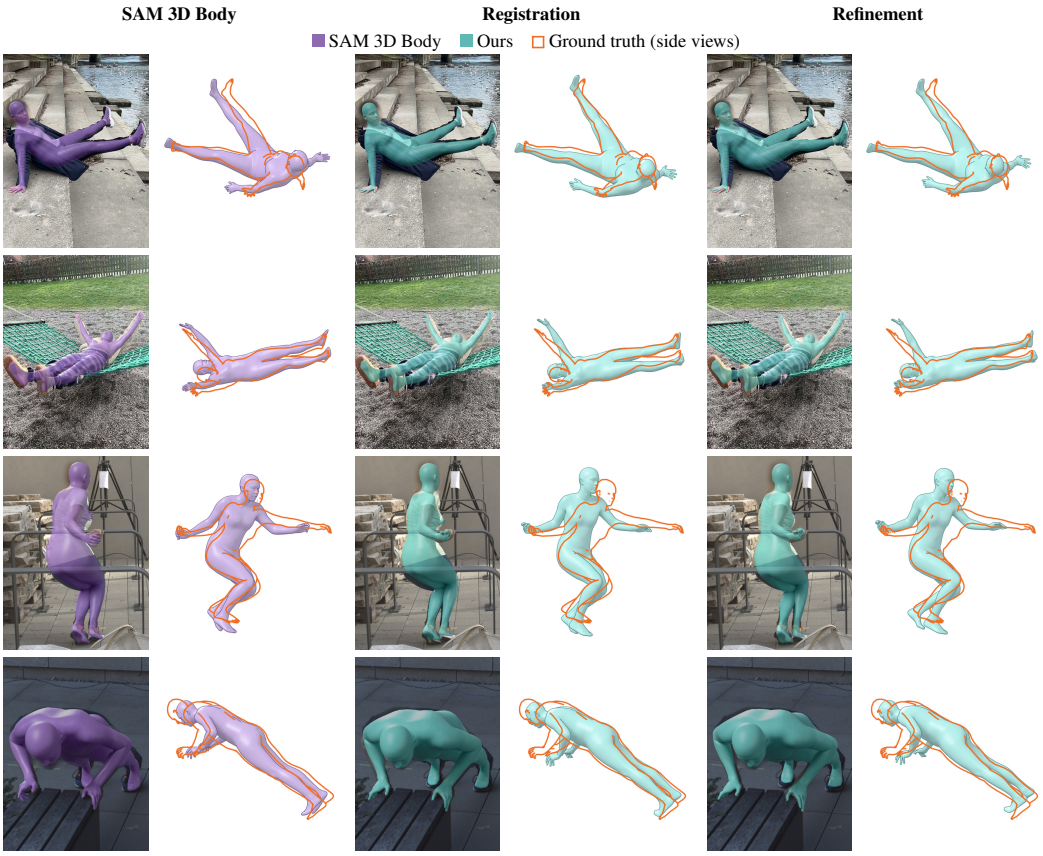

Figure 6: PixelBody's stages on the main paper's Figure 4 clips: SAM 3D Body initialization, registration, and refinement (AMASS-decoder model), over the input and from the side with the pelvis on the ground truth's (outline). Faces are blurred.

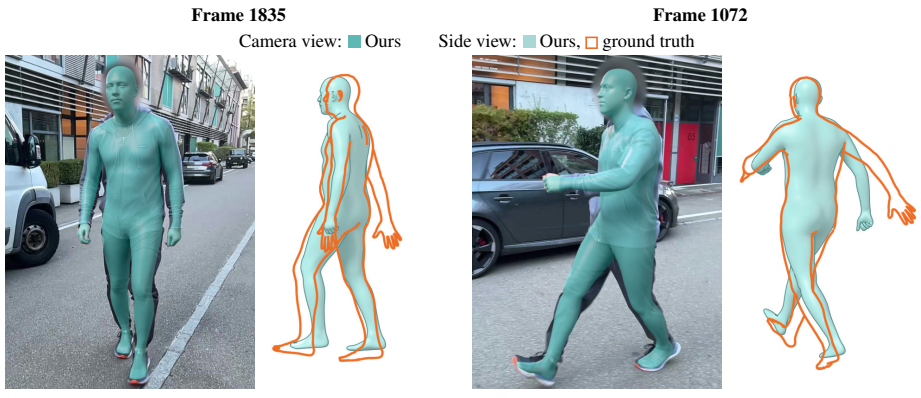

Figure 7: Failure cases of the AMASS-decoder model on two EMDB2 P3_28 frames chosen for high per-frame PA-MPJPE of an earlier solver version, each over the input frame and from the side with its pelvis on the ground truth's (outline): the arms swing too little in depth. Faces are blurred.