HER: HUMAN–ENVIRONMENT RECONSTRUCTION VIA VISUAL GEOMETRY-MEDIATED REASONING

Anonymous authors
Paper under double-blind review

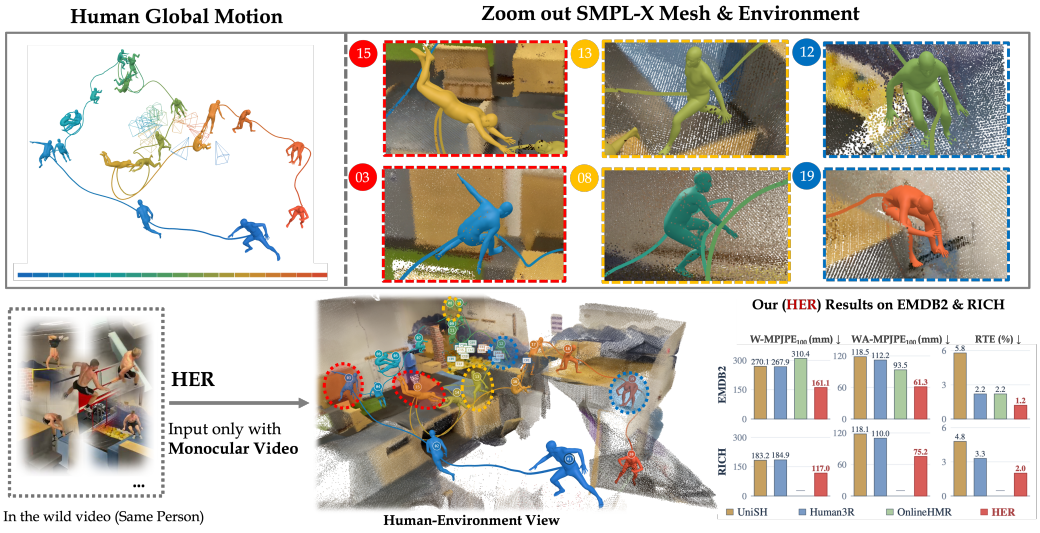


Figure 1: **Human-Environment Reconstruction (HER) from monocular video.** HER recovers human motion, camera trajectories, and scene geometry in a shared metric world frame. We show the global reconstruction, close-ups of human-environment results, and quantitative comparisons on EMDB2 and RICH.

ABSTRACT

Recent monocular HMR methods have made strong progress in world-space motion recovery, yet robust global trajectory estimation remains challenging under monocular ambiguity. Feedforward 3D reconstruction offers complementary spatial cues by recovering camera motion and the geometry of the moving human. We introduce HER, which leverages the moving-human geometry recovered by feedforward 3D reconstruction as a spatial bridge between parametric body prediction and the reconstructed world. HER builds on a backbone-agnostic feedforward geometry frontend to recover spatially consistent geometry and camera trajectories over long videos. For human reconstruction, we freeze a pretrained image HMR encoder and learn a temporal human representation that jointly supports SMPL-X body and explicit structural joint prediction. We further learn a geometry-mediated spatial reasoner that couples this representation with reconstructed human geometry to recover temporally consistent global translation while preserving high-quality pose and shape. Human body structure provides an additional scale reference for establishing a common world coordinate system. A lightweight world-space optimizer further improves temporal and contact consistency. Experiments on EMDB2 and RICH show that HER outperforms state-of-the-art methods in world-space human reconstruction while maintaining strong local accuracy, with robust performance on challenging in-the-wild videos.

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.

1 INTRODUCTION

Monocular human mesh recovery (HMR) has progressed from estimating camera-space pose and shape toward recovering human motion in world coordinates (Shin et al., 2024; Shen et al., 2024; Wang et al., 2024b). Yet reliable global trajectories remain difficult to recover because monocular observations leave ambiguities in depth, scale, and the separation of human and camera motion. Accurate local articulation therefore does not necessarily imply accurate placement in the surrounding environment. Such spatial grounding is also valuable for downstream motion retargeting and humanoid learning, where body-environment relationships matter (Cao et al., 2026; Yang et al., 2025; Allshire et al., 2025).

Our central observation is that reconstructed human geometry should not be treated merely as a dynamic outlier to be removed from scene reconstruction. Although this geometry is not an articulated human representation, it directly provides spatial evidence of where the observed person lies in the reconstruction frame. Conversely, parametric HMR provides articulated pose and metric body structure, but only weak global spatial grounding. HER exploits this complementarity by using reconstructed human geometry as a spatial bridge between parametric human prediction and the reconstructed world. Visual geometry drives spatial reasoning, while HER's human branch predicts SMPL-X pose and shape (Pavlakos et al., 2019) together with a separate root-relative metric skeleton for scale reasoning.

At the core of *Visual Geometry-Mediated Reasoning* (VGMR), joint queries initialized from global point features read encoded dynamic human geometry through attention. Their anatomical roles organize spatial evidence without requiring fixed joint-to-surface correspondences. The resulting spatial representation feeds two parallel branches. Body-structure scale reasoning combines the metric skeleton with the spatial joint features before temporal refinement to estimate the conversion from reconstruction units to body-model units. Asymmetric joint-centric temporal reasoning exchanges information across time only between corresponding joint tokens, then propagates this evidence to translation through within-frame interaction. Together, these branches recover camera-space human translation while retaining the pose and shape supplied by the human branch.

To establish the geometric reference, HER uses a backbone-agnostic front-end built on a VGGT-SLAM-style reconstruction framework (Maggio et al., 2025). Cross-window associations and verified loop constraints support scale synchronization and global camera optimization. Human regions are excluded from static-world alignment, while their reconstructed points remain available to VGMR. The human-derived scale predictions support per-frame human translation and, through sequence-level aggregation, set a common metric scale for camera translations and scene geometry. HER then places the human in the world frame and refines only its translation using reprojection, temporal, and geometrically valid contact constraints, keeping body pose, shape, cameras, scene geometry, and recovered scales fixed.

Experiments on EMDB2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022) show that HER obtains the lowest value of W-MPJPE$_{100}$ and WA-MPJPE$_{100}$ among the published results listed under the settings shown in 1. Camera evaluations and ablations assess the geometric front-end and the contributions of the human reasoning components. Qualitative results illustrate human–environment reconstruction in the wild.

Our contributions are threefold:

- We introduce dynamic human geometry as a spatial bridge between parametric human prediction and feedforward scene reconstruction. Unlike static-scene geometry used primarily for camera estimation, reconstructed human points directly provide person-centric spatial evidence in the reconstruction frame.
- We develop Visual Geometry-Mediated Reasoning (VGMR), which uses anatomically indexed queries to extract spatial evidence from reconstructed human geometry, combines it with a metric skeletal reference for scale reasoning, and propagates joint-centric temporal evidence to global translation.
- We integrate this reasoning into a metric world-space reconstruction system with scale-synchronized windowed geometry and translation-only refinement, achieving strong world-space human reconstruction while preserving local pose and shape.

2 Related Work

2.1 Human Mesh Recovery in Camera and World Coordinates

Camera-space HMR methods recover detailed body pose and shape from images (Kanazawa et al., 2018; Kolotouros et al., 2019; Wang et al., 2025c). Extending these predictions to world coordinates requires separating human motion from camera motion and resolving scale ambiguity. Existing methods combine motion priors with camera estimates, scene-derived scale, or contact constraints (Shin et al., 2024; Shen et al., 2024; Wang et al., 2024b; Zhao et al., 2026), while diffusion-based approaches further model plausible global motion (Li et al., 2025a; Wang et al., 2026c).

Existing methods primarily infer global motion from image features, camera estimates. In contrast, HER explicitly reasons over the geometry of the moving human, as a spatial observation rather than treating dynamic regions only as nuisances.

2.2 Visual Geometry Reconstruction and SLAM

DROID-SLAM combines learned updates with dense bundle adjustment (Teed & Deng, 2021). Feedforward models, including DUSt3R, VGGT, VGGT-Ω, and π^3, directly predict dense geometry (Wang et al., 2024a; 2025a; 2026a; 2025b). MASt3R-SLAM and VGGT-SLAM 1.0/2.0 integrate geometric predictions with global optimization (Murai et al., 2025; Maggio et al., 2025; Maggio & Carlone, 2026), while TTT3R adapts recurrent memory updates for sequential reconstruction (Chen et al., 2026a). MonST3R, MegaSaM, and Pi3MOS-SLAM further address geometry and camera estimation in dynamic scenes (Zhang et al., 2025; Li et al., 2025b; Zhong et al., 2025).

Existing dynamic-scene reconstruction methods primarily aim to protect camera and scene estimation from non-rigid motion. HER instead makes a dual use of dynamics: human regions are excluded from static alignment, but their reconstructed geometry is retained as input for human localization.

2.3 Joint Human-Environment Reconstruction

Joint human-environment reconstruction uses geometry and human structure to recover coherent bodies and environments. JOSH jointly optimizes humans, scenes, and cameras through contact constraints, while JOSH3R learns feedforward reconstruction from its pseudo-labels (Liu et al., 2025). Human3R jointly predicts bodies, scenes, and cameras (Chen et al., 2026b); UniSH supervises mesh–point-cloud correspondence (Li et al., 2026), and SHOW couples human scale priors with scene geometry (Shi et al., 2026). UniCon3R further uses inferred contact to correct human pose during inference (Sur et al., 2026). Complementary approaches include OnlineHMR's integration of causal human prediction with incremental SLAM (Zhao et al., 2026) and EmbodMocap's use of captured human–scene data to improve monocular reconstruction (Wang et al., 2026b).

HER connects human prediction with the reconstructed environment reference through geometry-mediated spatial and temporal reasoning. Human-structured queries extract evidence from dynamic human geometry, while a metric skeletal reference establishes the scale relation. The final refinement optimizes only world-space human translation, keeping pose, shape, cameras, scene geometry, and recovered scales fixed.

3 Method

3.1 Problem Formulation and Framework Overview

Given a monocular RGB video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, HER reconstructs SMPL-X human motion, camera trajectories, and scene geometry in a shared world coordinate system. As shown in Figure 2, the framework proceeds in three stages. First, a windowed visual geometry front-end estimates camera poses and point maps in a common reconstruction frame. Second, the human branch predicts body pose, shape, and a metric skeleton, while Visual Geometry-Mediated Reasoning (VGMR) uses reconstructed human geometry to estimate camera-space translation and the skeletal reference to infer scale. These estimates place the human, cameras, and scene in a common metric world frame. Finally,

a world-space optimizer refines human translation while keeping all other reconstructed quantities fixed.

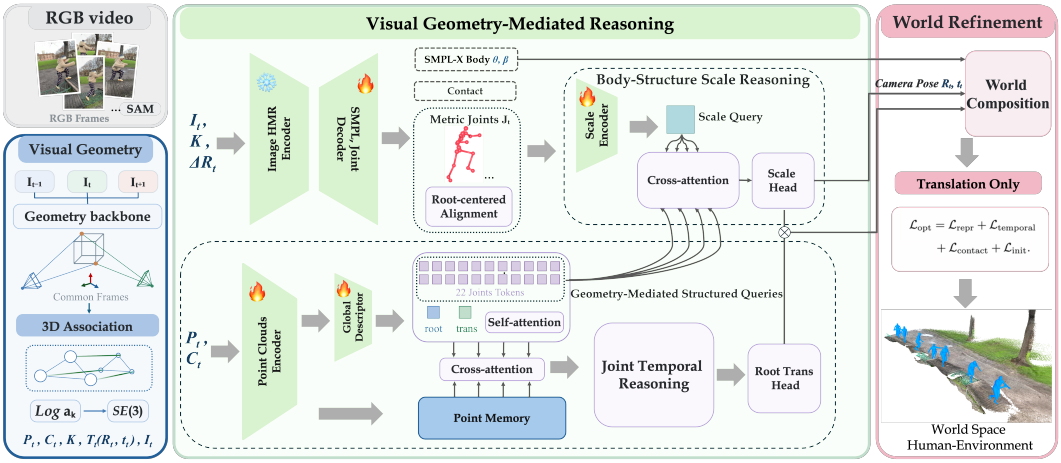


Figure 2: **HER Overview.** Geometry-initialized joint queries read spatial evidence from human point clouds, feeding parallel scale and temporal reasoning: a metric skeletal reference guides scale estimation, while joint-centric temporal reasoning supports human translation. Pose and shape are retained, followed by world-space translation refinement.

We first describe geometry-mediated human reconstruction in Section 3.2, taking the camera poses and point maps as inputs. We then present the visual geometry front-end in Section 3.3 and world-space translation refinement in Section 3.4.

3.2 Visual Geometry-Mediated Human Reconstruction

3.2.1 Human Pose and Skeletal Reference

For each tracked person, the image branch takes the frame, a SAM2 (Ravi et al., 2024) mask, its bounding box, and camera intrinsics. We freeze the PromptHMR image encoder and adapt its pretrained video head (Wang et al., 2025c), whose temporal design follows GVHMR (Shen et al., 2024). The head combines image features, 2D keypoints estimated by ViTPose++ (Xu et al., 2024), and camera information, including inter-frame rotation. The head predicts SMPL-X pose $\boldsymbol{\theta}_t$ and shape $\boldsymbol{\beta}$ (Pavlakos et al., 2019), together with a root-relative metric skeleton $\mathbf{J}_t \in \mathbb{R}^{22\times 3}$ constructed from predicted bone lengths and directions:

$$\mathbf{J}_{t,0} = \mathbf{0}, \qquad \mathbf{J}_{t,i} = \mathbf{J}_{t,\mathrm{parent}(i)} + \ell_i \widehat{\mathbf{d}}_{t,i}, \quad i = 1, \ldots, 21. \tag{1}$$

Here ℓ_i is the bone length shared within the temporal window, and $\widehat{\mathbf{d}}_{t,i}$ is the unit direction from the parent to joint i in camera coordinates. Predicted directly without SMPL-X mesh decoding, $\mathbf{J}_t$ provides the body-size reference for scale estimation. Implementation details appear in Section 4.2.

3.2.2 Geometry-Mediated Structured Queries

Let $\mathbf{P}_t$ denote the point map in the global reconstruction frame and $\mathbf{C}_t$ its confidence map. We select human pixels $\mathbf{u}_{t,n}$ within the tracked mask, discarding invalid points and low-confidence predictions according to $\mathbf{C}_t$. Using the camera-to-world pose $(\mathbf{R}_t, \mathbf{t}_t)$, we express each selected point in camera coordinates as $\mathbf{x}_{t,n} = \mathbf{R}_t^\top (\mathbf{P}_t(\mathbf{u}_{t,n}) - \mathbf{t}_t)$, retaining reconstruction units. The resulting human point set and its centroid are

$$\mathcal{P}_t^h = \{(\mathbf{x}_{t,n}, \mathbf{c}_{t,n}, \mathbf{n}_{t,n})\}_{n=1}^{N_t}, \qquad \boldsymbol{\mu}_t = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathbf{x}_{t,n}, \tag{2}$$

where $\mathbf{n}_{t,n}$ is the surface normal in camera coordinates, estimated from neighboring valid points, implementation details are provided in Appendix A.3. We center the coordinates as $\bar{\mathbf{x}}_{t,n} = \mathbf{x}_{t,n} - \boldsymbol{\mu}_t$, retaining the centroid for translation recovery. Utonia (Zhang et al., 2026) encodes the RGB-D

pointclouds. We fuse its features into a global descriptor $\mathbf{g}_t$ and construct point memory $\mathbf{Z}_t^{\mathrm{pc}}$ from sampled centered coordinates and projected features.

An MLP maps $\mathbf{g}_t$ to 22 queries indexed by the first 22 SMPL-X kinematic joints. We append two learned tokens shared across frames: $\mathbf{e}_{\mathrm{root}}$ for root bias and $\mathbf{e}_{\mathrm{trans}}$ for translation. All 24 tokens have width d. Each attention block applies query self-attention followed by cross-attention to the point memory:

$$\begin{aligned} \mathbf{Q}_t^{\mathrm{init}} &= [\mathrm{MLP}(\mathbf{g}_t); \mathbf{e}_{\mathrm{root}}; \mathbf{e}_{\mathrm{trans}}], \\ \mathbf{Q}_t &= \mathrm{QueryAttn}(\mathbf{Q}_t^{\mathrm{init}}, \mathbf{Z}_t^{\mathrm{pc}}). \end{aligned} \tag{3}$$

The updated tokens are denoted $\mathbf{Q}_t = [\mathbf{Q}_t^J; \mathbf{q}_t^{\mathrm{root}}; \mathbf{q}_t^{\mathrm{trans}}]$.

3.2.3 Parallel Scale and Temporal Modeling

The spatial tokens feed two parallel branches: one predicts scale using the metric skeleton, while the other updates the query features across time for translation recovery.

Body-structure scale reasoning. An MLP embeds the flattened metric skeleton $\mathrm{vec}(\mathbf{J}_t)$ as a scale query, which attends to the spatial joint tokens $\mathbf{Q}_t^J$ before temporal refinement. From the resulting feature $\mathbf{z}_t$, the head g_s predicts log scale γ_t, yielding a positive conversion factor s_t from reconstruction units to metric body-model units:

$$\begin{aligned} \mathbf{z}_t &= \mathrm{CrossAttn}\big(\mathrm{MLP}(\mathrm{vec}(\mathbf{J}_t)), \mathbf{Q}_t^J, \mathbf{Q}_t^J\big), \\ \gamma_t &= g_s(\mathbf{z}_t), \\ s_t &= e^{\gamma_t}. \end{aligned} \tag{4}$$

Asymmetric joint-centric temporal reasoning. Within each L-frame window, we apply temporal self-attention to tokens with the same joint index, using RoPE to encode frame positions. The updated joint tokens then interact with the root-bias and translation tokens through within-frame self-attention:

$$\begin{aligned} \widetilde{\mathbf{Q}}_{1:L,j}^J &= \mathrm{TempAttn}\left(\mathbf{Q}_{1:L,j}^J\right), \qquad j = 0, \ldots, 21, \\ \mathbf{Q}_t' &= \mathrm{FrameAttn}\left([\widetilde{\mathbf{Q}}_t^J; \mathbf{q}_t^{\mathrm{root}}; \mathbf{q}_t^{\mathrm{trans}}]\right), \qquad t = 1, \ldots, L, \end{aligned} \tag{5}$$

Here $\mathbf{Q}_{1:L,j}^J$ denotes the token sequence for joint j.

3.2.4 Translation Recovery and World-Space Composition

The translation head g_τ maps $\mathbf{q}_t^{\mathrm{trans}\prime}$ to a centroid-relative SMPL-X translation offset in reconstruction units. The root-bias output restores geometry-decoded joint positions and is not added to the translation offset.

From the raw scale predictions s_t, we obtain temporally smoothed per-frame scales s_t^{e} for human translation and a robust sequence-level scale s_g for camera translations and scene geometry. The metric camera-space SMPL-X translation is

$$\boldsymbol{\tau}_t^c = s_t^{\mathrm{e}}\big(\boldsymbol{\mu}_t + g_\tau(\mathbf{q}_t^{\mathrm{trans}\prime})\big). \tag{6}$$

World-space composition. For a camera-space human point $\mathbf{X}_t^c$ already expressed in metric units, its world-space position is

$$\mathbf{X}_t^w = \mathbf{R}_t \mathbf{X}_t^c + s_g \mathbf{t}_t. \tag{7}$$

The transformed motion initializes translation refinement (Section 3.4), with local articulation and body shape unchanged. The SMPL-X translation convention is detailed in Appendix A.2.

3.3 Visual Geometry Front-End

The visual frontend is not the primary learning contribution of HER. Its role is to provide VGMR with a spatially consistent reconstruction frame over long videos. We therefore adopt a VGGT-SLAM-style (Maggio et al., 2025) windowed design and introduce the scale synchronization and dynamic-region handling needed by our downstream human reasoning. It operates on geometry and camera predictions, allowing different feedforward backbones.

3.3.1 FEEDFORWARD GEOMETRY AND CROSS-WINDOW ASSOCIATION

For each overlapping reconstruction window $\mathcal{W}_k$, the backbone predicts point maps $\mathbf{P}_{k,t}$, confidence maps $\mathbf{C}_{k,t}$, intrinsics $\mathbf{K}_{k,t}$, and camera-to-window poses $\mathbf{T}_{k,t}$ for $t \in \mathcal{W}_k$. The point maps are expressed in the local window frame. Our default backbone is Pi3X (Wang et al., 2025b).[1]

We match image pairs from neighboring windows and use SALAD (Izquierdo & Civera, 2024) to retrieve non-adjacent revisit candidates. For a pair from windows i and j, RoMaV2 (Edstedt et al., 2026) provides pixel correspondences $(\mathbf{u}_n, \mathbf{v}_n)$, lifted to 3D as $\mathbf{p}_n = \mathbf{P}_{i,t}(\mathbf{u}_n)$ and $\mathbf{q}_n = \mathbf{P}_{j,t'}(\mathbf{v}_n)$. Valid correspondences outside human masks form static-scene constraints, while reconstructed human points are retained for VGMR.

3.3.2 ROBUST CROSS-WINDOW SCALE SYNCHRONIZATION

Let $a_k > 0$ map window k to common reconstruction units. Since pairwise distances are invariant to rigid transformations, two lifted correspondences m and n provide a relative-scale observation independently of window rotation and translation:

$$r_{ij}^{(m,n)} = \frac{\|\mathbf{p}_m - \mathbf{p}_n\|_2}{\|\mathbf{q}_m - \mathbf{q}_n\|_2}, \qquad r_{ij} = \mathrm{RobustAggregate}_{(m,n)\in\mathcal{D}_{ij}}\, r_{ij}^{(m,n)}, \tag{8}$$

Here r_{ij} robustly aggregates observations over valid, non-degenerate pairs $\mathcal{D}_{ij}$ and estimates a_j/a_i.

We optimize log scales over accepted neighboring and long-range edges $\mathcal{E}_s$:

$$\begin{aligned} \min_{\substack{a_k>0 \\ a_0=1}} \quad & \sum_{(i,j)\in\mathcal{E}_s} \rho_s\Big(w_{ij}\left(\log(a_j/a_i) - \log r_{ij}\right)^2\Big) \\ & + \mathcal{L}_{\mathrm{smooth}} + \mathcal{L}_{\mathrm{prior}}. \end{aligned} \tag{9}$$

Here w_{ij} combines a source-dependent base weight with geometric reliability, and ρ_s is the soft-L1 penalty. The regularizers impose second-order smoothness in log scale and a weak prior toward unit correction. Fixing $a_0 = 1$ selects the reconstruction-scale reference; the human-derived s_g subsequently converts it to metric units.

3.3.3 GLOBAL WINDOW-POSE OPTIMIZATION

We rescale local points and camera translations by a_k, then hold these scales fixed. The pose factors $\mathcal{E} = \mathcal{E}_{\mathrm{ov}} \cup \mathcal{E}_{\mathrm{align}} \cup \mathcal{E}_{\mathrm{loop}}$ come from shared-frame overlap, rigid alignment of rescaled RoMaV2 correspondences, and geometrically verified SALAD loop candidates, respectively. Scale and pose constraints are verified separately; retrieval alone adds neither. We optimize rigid window-to-world transforms $\mathbf{G}_k \in SE(3)$:

$$\{\mathbf{G}_k^\star\} = \arg\min_{\{\mathbf{G}_k\}} \sum_{e\in\mathcal{E}} \rho_e\Big(\big\|\mathrm{Log}\big(\mathbf{Z}_e^{-1}\mathbf{G}_i^{-1}\mathbf{G}_j\big)\big\|_{\Omega_e}^2\Big). \tag{10}$$

For factor e between windows i and j, $\mathbf{Z}_e$ maps the scale-normalized frame of j to that of i, Ω_e is its information matrix, and ρ_e is a robust penalty. The map Log gives a six-dimensional pose residual; one reference window pose is fixed. Applying the optimized transforms aligns camera poses and point maps, including the dynamic human points used by VGMR. Further details appear in Appendix A.3.

3.4 WORLD-SPACE TRANSLATION REFINEMENT

We refine the world-space motion from Section 3.2.4 through translation alone, keeping body pose, shape, cameras, scene geometry, and scales fixed. A per-frame residual $\boldsymbol{\delta}_t^w \in \mathbb{R}^3$ corrects the initial SMPL-X translation $\boldsymbol{\tau}_t^w$:

$$\boldsymbol{\tau}_t^{w,\mathrm{ref}} = \boldsymbol{\tau}_t^w + \boldsymbol{\delta}_t^w. \tag{11}$$

The residuals $\{\boldsymbol{\delta}_t^w\}_{t=1}^T$ minimize

$$\mathcal{L}_{\mathrm{opt}} = \lambda_{\mathrm{repr}}\mathcal{L}_{\mathrm{repr}} + \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}} + \lambda_{\mathrm{contact}}\mathcal{L}_{\mathrm{contact}} + \lambda_{\mathrm{init}}\mathcal{L}_{\mathrm{init}}. \tag{12}$$

[1] Pi3X is an engineering update to π^3: `https://github.com/yyfz/Pi3`.

Here $\mathcal{L}_{\text{repr}}$ penalizes reprojection errors between SMPL-X joints and observed 2D keypoints, and $\mathcal{L}_{\text{temp}}$ encourages temporal coherence. The contact loss $\mathcal{L}_{\text{contact}}$ is applied only to predicted support contacts that pass a geometric consistency check under a static-world assumption, while $\mathcal{L}_{\text{init}}$ keeps the refined translation close to its initial estimate. The body update and reprojection formulations are given in Appendix A.4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate world-space human reconstruction on EMDB2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022), reporting PA-MPJPE for aligned pose, W-MPJPE$_{100}$ and WA-MPJPE$_{100}$ under their respective 100-frame protocols, and RTE for root trajectories. Human ablations use EMDB2 with shared inputs, geometry, and evaluator settings, and additionally report jitter and foot skating; backbone and scale studies use two-epoch pilot training. Camera absolute trajectory error (ATE) RMSE is evaluated on 25 EMDB2 sequences in meters and eight Bonn RGB-D sequences (Palazzolo et al., 2019) in centimeters. Lower is better for all metrics. Baseline scores come from original papers and public results, with camera settings indicated in the tables. Appendix B records published evaluation conventions and comparison sources.

4.2 IMPLEMENTATION DETAILS

We trained VGMR on BEDLAM1 (Black et al., 2023) and the video head on BEDLAM1 and 3DPW (von Marcard et al., 2018). The PromptHMR image encoder is frozen, while the video head is retrained for SMPL-X and metric skeletal prediction (Appendix A.1.1). From the same raw predictions s_t, we obtain the per-frame human scale s_t^{c} by temporal smoothing and the global scene scale s_g by robust aggregation. The former scales camera-space human translation; the latter uniformly scales camera translations and scene geometry. Additional implementation details are given in Appendix A.

4.3 MAIN RESULTS

4.3.1 WORLD-SPACE HUMAN RECONSTRUCTION

Table 1 compares human-only and joint human–scene methods on EMDB2 and RICH.

HER achieves the lowest reported errors across all four metrics on EMDB2 and the lowest WA-MPJPE$_{100}$ and W-MPJPE$_{100}$ on RICH. On RICH, these joint errors improve over UniCon3R by 7.7% and 9.9%, respectively, although RTE remains higher (2.0 versus 1.7).

4.3.2 CAMERA TRAJECTORY ACCURACY

Table 2 compares camera ATE and HER variants using different geometry backbones or removing dynamic masking. Uncalibrated denotes evaluation without supplied camera calibration.

HER achieves the lowest reported ATE, 0.30 m, compared with 0.47 m for Pi3MOS-SLAM under the listed calibration settings. Removing dynamic masking increases HER's error to 0.48 m. Replacing the backbone with VGGT or VGGT-Ω increases ATE to 2.07 m or 0.42 m, respectively, highlighting the contributions of dynamic filtering and local geometry quality.

Bonn RGB-D benchmark. HER achieves the lowest ATE on five of eight Bonn sequences. Without supplied intrinsics or uncertainty-based dynamic filtering, it remains competitive with Pi3MOS-SLAM (mean ATE: 2.32 vs. 2.20 cm; Table 8).

4.3.3 QUALITATIVE HUMAN–ENVIRONMENT RECONSTRUCTION

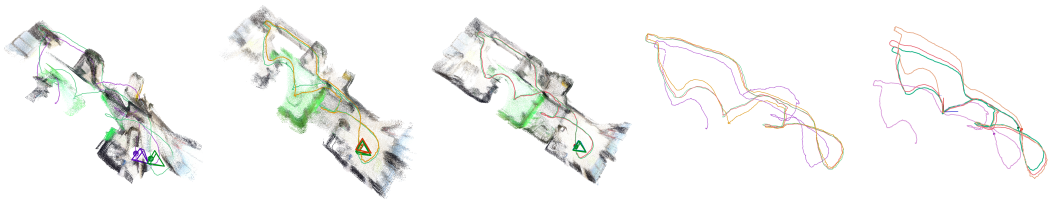
Figure 3 shows that HER recovers more stable trajectories and more accurate scale than OnlineHMR and Human3R on a long EMDB2 sequence.

Table 1: **Human global trajectory and motions.** Env. indicates environment reconstruction; dashes denote unreported values. Bold and underlined values mark the best and second-best results among methods with environment reconstruction.

		EMDB2				RICH			
Method	Env.	PA-MPJPE	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE	PA-MPJPE	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE
WHAM† (Shin et al., 2024)	✗	38.2	135.6	354.8	6.0	44.3	108.4	190.1	4.5
GVHMR† (Shen et al., 2024)	✗	–	111.0	276.5	2.0	39.5	78.8	126.3	2.4
TRAM (Wang et al., 2024b)	✗	38.1	76.4	222.4	1.4	–	127.8	238.0	6.0
PromptHMR-Vid (Wang et al., 2025c)	✗	–	71.0	216.5	1.3	37.0	–	–	–
JOSH (Liu et al., 2025)	✓	–	<u>68.9</u>	<u>174.7</u>	<u>1.3</u>	–	89.0	132.5	3.0
JOSH3R (Liu et al., 2025)	✓	–	220.0	661.7	13.1	–	190.4	334.9	6.3
Human3R (Chen et al., 2026b)	✓	–	112.2	267.9	2.2	–	110.0	184.9	3.3
UniSH (Li et al., 2026)	✓	44.4	118.5	270.1	5.8	50.0	118.1	183.2	4.8
SHOW (Shi et al., 2026)	✓	–	109.1	262.3	2.1	–	107.3	172.7	2.2
UniCon3R (Sur et al., 2026)	✓	<u>38.2</u>	113.7	285.6	2.3	<u>37.6</u>	<u>81.5</u>	<u>129.8</u>	**1.7**
OnlineHMR (Zhao et al., 2026)	✓	40.1	93.5	310.4	2.2	–	–	–	–
EmbodMocap (Wang et al., 2026b)	✓	–	82.2	220.7	1.7	–	–	–	–
Ours	✓	**31.4**	**61.3**	**161.1**	**1.2**	**37.1**	**75.2**	**117.0**	<u>2.0</u>

Table 2: **Camera ATE RMSE (m) ↓ on EMDB2.** *The SLAM pipeline adopted in OnlineHMR (Zhao et al., 2026); †uncertainty-based dynamic filtering.

Method	w Det.	w/o Calib.	ATE RMSE (m) ↓
VGGT-SLAM 1.0 (Maggio et al., 2025)	✗	✓	7.68
VGGT-SLAM 2.0 (Maggio & Carlone, 2026)	✗	✓	2.95
MASt3R-SLAM (Murai et al., 2025)	✗	✗	1.24
MASt3R-SLAM* (Murai et al., 2025)	✓	✗	0.98
MegaSaM (Li et al., 2025b)	✓†	✓	0.84
Pi3MOS-SLAM (Zhong et al., 2025)	✓†	✗	0.47
Ablations of our method			
Ours (w/o masking)	✗	✓	0.48
Ours (VGGT) (Wang et al., 2025a)	✓	✓	2.07
Ours (VGGT-Ω) (Wang et al., 2026a)	✓	✓	0.42
Full method			
Ours	✓	✓	**0.30**



(a) Human3R (b) OnlineHMR (c) HER (Ours) (d) Sim(3) aligned (e) First frame

Figure 3: **World-space reconstruction comparison.** (a–c) Human3R, OnlineHMR, and HER. Trajectories use (d) Sim(3) alignment or (e) a unit-scale rigid transform fitted to first-frame predicted/GT 3D joints and applied throughout. Red, orange, pink, and green denote HER, OnlineHMR, Human3R, and GT, respectively.

4.4 Ablation Studies

4.4.1 Geometry-Mediated Network Ablation

Table 3 isolates localization, skeletal reference, video head, temporal reasoning, and final refinement on EMDB2.

Table 3: **HER ablation on EMDB2.** Joints compares direct metric and SMPL-X-regressed joints; Head compares our adapted and original pretrained PromptHMR video heads. Points Align uses root-to-point median localization; * replaces only its scale with ground truth.

Mode	VGMR	Joints	Head	Temp.	Post-Proc.	W-MPJPE$_{100}$ ↓	WA-MPJPE$_{100}$ ↓	RTE ↓	Jitter ↓	Foot Skate ↓
VGMR Off	✗	–	✓	–	✓	341.5	120.2	3.51	13.6	7.5
VGMR Off (Points Align)	✗	–	✓	–	✓	173.5	65.7	1.18	<u>12.7</u>	5.6
VGMR Off*(GT Scale)	✗	–	✓	–	✓	173.1	64.6	**0.57**	<u>12.7</u>	5.6
SMPL-X Joints	✓	✗	✓	✓	✓	<u>163.8</u>	63.1	<u>1.15</u>	14.6	**5.3**
PHMR-Vid Head	✓	✗	✗	✓	✓	168.2	63.5	1.21	15.6	**5.3**
Temp. Off	✓	✓	✓	✗	✓	<u>163.8</u>	<u>62.6</u>	1.17	<u>12.7</u>	<u>5.4</u>
Post-Proc. Off	✓	✓	✓	✓	✗	167.4	66.1	1.16	129.9	17.1
HER	✓	✓	✓	✓	✓	**161.1**	**61.3**	<u>1.15</u>	**12.6**	**5.3**

Using the scale obtained by aligning estimated and ground-truth camera poses reduces Points Align's RTE from 1.18 to 0.57, highlighting sensitivity to metric-scale error. HER retains lower joint errors. Appendix C.3 provides the analysis.

4.4.2 Point-Cloud Backbone and Scale Reasoning

Two-epoch studies on EMDB2 compare point-cloud backbones and scale reasoning.

Backbone architecture. Comparison between PTv3 (Wu et al., 2024), LitePT (Yue et al., 2026), and Utonia (Zhang et al., 2026) can be found in Appendix C.5. Utonia achieves the best performance across all three metrics, supporting its use as our point-cloud backbone.

Scale reasoning. Table 4 shows that Body-Structure Scale Reasoning yields lower joint errors than explicit joint fitting and direct scale regression, while matching the best RTE.

Table 4: **Scale-estimation ablation.** Two-epoch results on EMDB2. Joint indicates use of a skeletal reference for scale estimation. Bold denotes the lowest reported value, including ties.

Scale estimation	Joint	W-MPJPE$_{100}$ ↓	WA-MPJPE$_{100}$ ↓	RTE ↓
Explicit joint fitting	✓	182.2	72.7	1.42
Direct scale regression	✗	181.1	72.4	**1.29**
Ours (Body-Structure Reasoning)	✓	**176.8**	**71.8**	**1.29**

5 Conclusion

HER leverages reconstructed human geometry as a spatial bridge between human representations and the surrounding 3D world, enabling global localization and metric-scale reconstruction while preserving local pose and shape. Experiments on EMDB2 and RICH demonstrate strong world-space reconstruction accuracy. More broadly, our results suggest that dynamic human geometry need not be treated solely as an outlier in geometric reconstruction. When coupled with articulated human priors and metric body structure, it can serve as a useful spatial observation for grounding human motion in a shared reconstructed world.

AI USE STATEMENT

In this work, we used Fable 5 (Anthropic) solely for language polishing to improve grammar and fluency. Generative AI tools were not used for generating synthetic data, developing theoretical frameworks, conducting statistical analyses, interpreting results, or drawing research conclusions. All AI-assisted edits were manually verified by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work studies human–environment reconstruction from monocular videos using existing datasets. We do not collect new human-subject data. Potential privacy concerns may arise when applying human reconstruction to videos without appropriate consent, and we encourage responsible use of the method.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model, training, implementation, and evaluation settings in the main paper and appendix. The source code and trained models will be made publicly available.

REFERENCES

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2025.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, June 2023.

Hanyang Cao, Yuetong Fang, Taesoo Kwon, Runyi Yu, Ji Ma, Jing Tan, Yangchen Zhou, Baoze Du, Yi Gu, Yukang Gao, et al. Unified motion retargeting for humanoids with learned point cloud correspondence. *arXiv preprint arXiv:2609.02134*, 2026.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D Reconstruction as Test-Time Training. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://arxiv.org/abs/2509.26645.

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)*, 2026b. URL https://arxiv.org/abs/2510.06219.

Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. RoMa v2: Harder Better Faster Denser Feature Matching. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2511.15706.

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.

Sergio Izquierdo and Javier Civera. Optimal Transport Aggregation for Visual Place Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17658–17668, 2024.

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 conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14632–14643, 2023.

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

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

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. *arXiv preprint arXiv:2601.01222*, 2026. URL https://arxiv.org/abs/2601.01222.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10486–10496, 2025b.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint Optimization for 4D Human-Scene Reconstruction in the Wild. *arXiv preprint arXiv:2501.02158*, 2025. URL https://arxiv.org/abs/2501.02158.

Dominic Maggio and Luca Carlone. VGGT-SLAM 2.0: Real-time Dense Feed-forward Scene Reconstruction. *arXiv preprint arXiv:2601.19887*, 2026. URL https://arxiv.org/abs/2601.19887.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.12549.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705, 2025.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019. URL https://www.ipb.uni-bonn.de/pdfs/palazzolo2019iros.pdf.

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)*, 2019.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

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 Conference Proceedings*, 2024. URL https://arxiv.org/abs/2409.06662.

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. URL https://arxiv.org/abs/2606.27720.

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.

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. URL https://arxiv.org/abs/2604.19923.

Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569, 2021.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. URL https://arxiv.org/abs/2605.15195.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024a.

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. *arXiv preprint arXiv:2602.23205*, 2026b. URL https://arxiv.org/abs/2602.23205.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025b. URL https://arxiv.org/abs/2507.13347.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos. *arXiv preprint arXiv:2403.17346*, 2024b. URL https://arxiv.org/abs/2403.17346.

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, 2025c.

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

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point Transformer V3: Simpler Faster Stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851, 2024.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer foundation model for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:1212–1230, 2024. doi: 10.1109/TPAMI.2023.3330016.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

Yuanwen Yue, Damien Robert, Jianyuan Wang, Sunghwan Hong, Jan Dirk Wegner, Christian Rupprecht, and Konrad Schindler. LitePT: Lighter Yet Stronger Point Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2512.13689.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.03825.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward One Encoder for All Point Clouds. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2603.03283.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and Laszlo 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)*, 2026. URL https://arxiv.org/abs/2603.17355.

Xingguang Zhong, Liren Jin, Marija Popović, Jens Behley, and Cyrill Stachniss. Dynamic Visual SLAM using a General 3D Prior. *arXiv preprint arXiv:2512.06868*, 2025. URL https://arxiv.org/abs/2512.06868.

A Additional Method Details

This section details the training stages, scale and coordinate conventions, SLAM constraints, and world-space translation refinement. Appendix B records evaluation conventions and comparison sources; Appendix C presents additional quantitative results and analysis; Appendix D provides qualitative reconstructions.

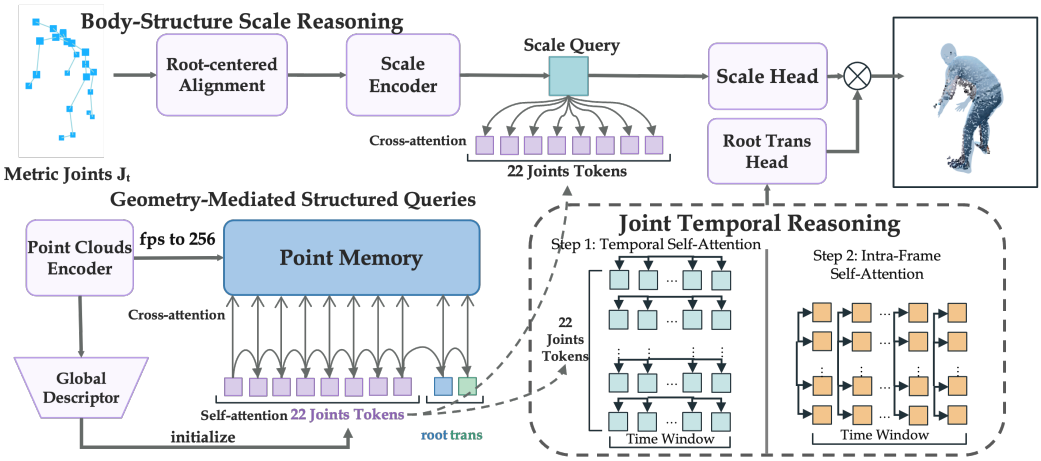


Figure 4: **Detailed pipeline of VGMR.** Geometry-initialized joint queries interact with point memory to produce joint tokens that feed two parallel branches: one combines with a scale query to predict scale, while the other refines these features over time to recover translation.

A.1 Training Details of Human Reconstruction

Table 5 summarizes the two training schedules and the fixed components used by each stage. The human image encoder remains frozen throughout video-head adaptation. The video-head and VGMR schedules have separate epoch counts; their rows describe the stages within each schedule.

Table 5: **Training stages and fixed inputs.** Video-head adaptation excludes pretraining. H1–H3 and G1–G3 index stages within their respective schedules. Epoch counts do not include the separate two-epoch pilot studies.

Stage	Adaptation target	Input or fixed component	Epochs
H1	Direct metric skeletal prediction	Pretrained video-head initialization; frozen image encoder	30
H2	Bone-length and direction branches	Frozen image encoder	10
H3	Skeletal-reference adaptation	Frozen image encoder and VGMR; BEDLAM1/3DPW mixture	10
G1	Spatial module	BEDLAM1 human point clouds from masked RGB-D	20
G2	Scale module	Spatial network fixed at its final checkpoint	10
G3	Temporal module and heads	Cached features from the frozen spatial network, without augmentation	50

A.1.1 HUMAN POSE AND SKELETAL REFERENCE

We adapt a pretrained video head on four NVIDIA RTX 5090 GPUs using stages H1–H3 in Table 5. The skeletal branch uses 23 unique observations from 25 detected keypoints to predict 22 SMPL-X kinematic joints, with additional foot-keypoint conditioning.

Stage H3 mixes BEDLAM1 and 3DPW at approximately 3:1 using 120-frame windows, AdamW (weight decay 0.01), and bfloat16. The skeletal prediction head and bone-length readout use a fixed learning rate of 3×10^{-5}. Each GPU processes 32 windows per microbatch; four accumulation steps give an effective batch of 512 windows. We use the final checkpoint after 720 updates, without validation-based checkpoint selection.

Stage H3 objectives and gradient routing. Skeletal supervision uses pelvis-relative joints in camera coordinates and meters:

$$\mathcal{L}_{\text{skel}} = \mathcal{L}_{\text{pos}} + 2\mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{acc}} + 0.1\mathcal{L}_{\text{len}}. \tag{13}$$

Position, first- and second-order temporal differences, and bone lengths use Smooth L1 thresholds $(0.02, 0.002, 0.001, 0.005)$ and divisors $(0.02, 0.01, 0.01, 0.02)$, respectively; direction loss averages $1 - \widehat{\mathbf{d}} \cdot \mathbf{d}^*$ for predicted and target unit bone directions. Only valid joints or consecutive valid frames contribute. An auxiliary camera-space pelvis loss through frozen VGMR uses Smooth L1 threshold 0.05, divisor 0.1, and weight 0.40358 with training-frame weights. Only $\mathcal{L}_{\text{len}}$ updates the bone-length branch; the remaining terms update the direction branch, with conflict projection applied to auxiliary gradients.

A.1.2 VGMR

VGMR uses 30 FPS BEDLAM1 human point clouds backprojected from masked RGB-D, with XYZ, RGB, and normal channels. We use 0.01 m voxel sampling and at most 2048 points per frame, without scale normalization. Stages G1–G3 train the spatial, scale, and temporal modules, respectively (Table 5). The spatial checkpoint is fixed for scale training and for feature caching in temporal training. Training takes approximately three days on two NVIDIA H100 GPUs with bfloat16 and an effective batch of 128.

AdamW uses weight decay 0.02, learning rates 10^{-4} for Utonia and 10^{-3} for other trainable modules, and gradient-norm clipping at 1.0. Cosine OneCycleLR uses `pct_start=0.05`, `div_factor=10`, and `final_div_factor=1000`.

Augmentation. Spatial and scale training always apply isotropic scaling in $[0.5, 2.0]$ and Gaussian point jitter ($\sigma = 0.01$, clipped to $[-0.05, 0.05]$). We also apply left–right flips with joint permutation ($p = 0.5$) and rotations about X, Y, or Z in $[-\pi, \pi]$ ($p = 0.5$). Scale training additionally jitters the reference skeleton with $\sigma = 0.01$ Gaussian noise.

Spatial module. Four-stage Utonia uses blocks $[3, 3, 3, 12]$, channels $[54, 108, 216, 432]$, patch size 1024, drop-path rate 0.3, and flash attention. PointNet++-style fusion module Qi et al. (2017) maps its 432-dimensional features to a 1024-dimensional global descriptor. The query decoder has two layers, two heads, dropout 0.1, and FFN width 256. The losses $\mathcal{L}_{\text{joint}}$, $\mathcal{L}_{\text{root}}$, and $\mathcal{L}_{\text{trans}}$ are coordinate-wise MSEs for the 22 root-relative joints, centroid-relative root bias, and centroid-relative SMPL-X translation, respectively:

$$\mathcal{L}_{\text{VGMR}} = \mathcal{L}_{\text{joint}} + 0.2\mathcal{L}_{\text{root}} + 0.2\mathcal{L}_{\text{trans}}. \tag{14}$$

Scale module. The scale module uses one cross-attention layer and predicts $s_t = e^{\gamma_t}$, with target $s_{\text{gt}} = 1/s_{\text{aug}}$ for augmentation factor s_{aug}. For spatial joint predictions $\hat{\mathbf{J}}_t$ and ground-truth joints $\mathbf{J}_t^{\text{gt}}$, its objective is

$$\begin{aligned} \mathcal{L}_{\log-\text{scale}} &= \|\gamma_t - \log(s_{\text{gt}})\|_2^2, \\ \mathcal{L}_{\text{recon}} &= \text{MSE}\Big(\text{stopgrad}(\hat{\mathbf{J}}_t)s_t, \mathbf{J}_t^{\text{gt}}/s_{\text{aug}}\Big), \\ \mathcal{L}_{\text{scale}} &= \mathcal{L}_{\log-\text{scale}} + 0.1\mathcal{L}_{\text{recon}}. \end{aligned} \tag{15}$$

Temporal module. Two temporal/frame-attention layers and dedicated prediction heads use 120-frame windows at 30 FPS with stride 5. Training reuses the three MSE terms in Eq. 14 with weights $(1, 1, 1)$.

A.2 Human Scale Aggregation and Coordinate Conventions

Table 6 distinguishes window-scale synchronization from metric conversion and records the coordinate frames of the quantities used below. Reconstruction units refer to the common SLAM reference after window corrections a_k; metric quantities use meters.

Table 6: **Coordinate frames and units.** The metric skeletal reference and geometry-decoded joints are separate predictions. Rotations change coordinate frames without changing units.

Quantity	Meaning and coordinate frame	Units or conversion
a_k	Window-scale correction inside SLAM	Local window units to common reconstruction units
$\mathbf{P}_t,\ \mathbf{t}_t$	Point map and camera center in the global reconstruction frame	Reconstruction units
$\mathbf{x}_{t,n},\ \boldsymbol{\mu}_t$	Human points and centroid in the camera frame	Reconstruction units
$\widehat{\mathbf{J}}^{\mathrm{rel}}_{t,j},\ \widehat{\mathbf{b}}_t$	Geometry-decoded root-relative joints and centroid-relative root bias, in camera axes	Reconstruction units
$g_\tau(\mathbf{q}_t^{\mathrm{trans}\prime})$	Centroid-relative SMPL-X translation offset, in camera axes	Reconstruction units
$\mathbf{J}_t$	Human-branch root-relative skeletal reference in camera axes	Meters
$s_t,\ s_t^{\mathrm{e}},\ s_g$	Raw, filtered human, and global scene scales	Meters per reconstruction unit
$\boldsymbol{\tau}_t^c,\ \boldsymbol{\tau}_t^w,\ \boldsymbol{\delta}_t^w$	Camera/world SMPL-X translation parameters and world translation correction	Meters

IQR-based scene scale. Let $\mathcal{T}_s$ contain frames with valid, finite, positive raw scale predictions s_t, pooled across inference segments. Let $q_{0.25}$ and $q_{0.75}$ be their lower and upper quartiles, and define $\mathrm{IQR} = q_{0.75} - q_{0.25}$. For multiplier $\kappa_{\mathrm{IQR}} \geq 0$, the inlier set and scene scale are

$$\begin{aligned} \mathcal{T}_{\mathrm{in}} &= \{t \in \mathcal{T}_s : q_{0.25} - \kappa_{\mathrm{IQR}}\,\mathrm{IQR} \leq s_t \leq q_{0.75} + \kappa_{\mathrm{IQR}}\,\mathrm{IQR}\}, \\ s_g &= \frac{1}{|\mathcal{T}_{\mathrm{in}}|} \sum_{t \in \mathcal{T}_{\mathrm{in}}} s_t, \qquad |\mathcal{T}_{\mathrm{in}}| > 0. \end{aligned} \tag{16}$$

The mean uses raw predictions, rather than the effective per-frame translation factors s_t^{e}.

Joint-root recovery. Let $\widehat{\mathbf{b}}_t$ be the root-bias head's centroid-relative offset and $\widehat{\mathbf{J}}^{\mathrm{rel}}_{t,j}$ the joint head's root-relative prediction, both in reconstruction units. The geometry branch recovers

$$\widehat{\mathbf{r}}_t = \boldsymbol{\mu}_t + \widehat{\mathbf{b}}_t, \qquad \widehat{\mathbf{J}}^{\mathrm{cam}}_{t,j} = \widehat{\mathbf{J}}^{\mathrm{rel}}_{t,j} + \widehat{\mathbf{r}}_t. \tag{17}$$

These geometry-decoded joints are distinct from the metric skeletal reference $\mathbf{J}_t$. The root-bias vector restores joint positions; it is not added to the separate SMPL-X translation prediction in Eq. 6.

Consistent units. The globally fused point map and camera-to-world translation share the same reconstruction units. For a human pixel $\mathbf{u}_{t,n}$, its camera-frame point is

$$\mathbf{x}_{t,n} = \mathbf{R}_t^\top \left(\mathbf{P}_t(\mathbf{u}_{t,n}) - \mathbf{t}_t\right). \tag{18}$$

The metric scene points and camera translations are $s_g\mathbf{P}_t$ and $s_g\mathbf{t}_t$. SMPL-X body offsets and $\boldsymbol{\tau}_t^c$ are already in meters and do not receive another factor of s_g. The window correction a_k has already been applied before these globally aligned quantities are passed to VGMR.

SMPL-X translation and pelvis position. Let $\mathbf{p}_\beta$ denote the shape-dependent pelvis in the untranslated SMPL-X model, which is also its global-rotation center. The SMPL-X parameter `transl` places the pelvis at $\mathbf{p}_\beta + \boldsymbol{\tau}_t^c$ in camera coordinates. Applying Eq. 7 to this position gives

$$\boldsymbol{\tau}_t^w = \mathbf{R}_t(\mathbf{p}_\beta + \boldsymbol{\tau}_t^c) + s_g\mathbf{t}_t - \mathbf{p}_\beta. \tag{19}$$

For a pelvis-relative body vertex $\mathbf{X}_{t,v}^{\mathrm{rel}}$ oriented in the camera frame, the corresponding world vertex is

$$\mathbf{X}_{t,v}^w = \mathbf{R}_t\mathbf{X}_{t,v}^{\mathrm{rel}} + \mathbf{p}_\beta + \boldsymbol{\tau}_t^w. \tag{20}$$

This is equivalent to transforming the camera-space vertex $\mathbf{X}_{t,v}^{\mathrm{rel}} + \mathbf{p}_\beta + \boldsymbol{\tau}_t^c$ directly. Equivalently, $\boldsymbol{\tau}_t^w = \mathbf{R}_t\boldsymbol{\tau}_t^c + s_g\mathbf{t}_t + (\mathbf{R}_t - \mathbf{I})\mathbf{p}_\beta$. The final term accounts for the SMPL-X rotation center: transforming the translation parameter alone would omit it. If $\mathbf{R}_t = \mathbf{I}$, this reduces to $\boldsymbol{\tau}_t^w = \boldsymbol{\tau}_t^c + s_g\mathbf{t}_t$. Body dimensions remain unchanged in both forms.

A.3 Visual Geometry SLAM Implementation

Inference-time normals. The human branch reads the organized point map produced by HERO-SLAM; it does not generate the map with an additional geometry model. Let $\mathbf{X}_t(r, c)$ be its camera-coordinate 3D point at pixel (r, c), for each interior pixel whose center and four immediate neighbors are valid, we compute

$$\mathbf{n}_{t,r,c} = \frac{[\mathbf{X}_t(r, c+1) - \mathbf{X}_t(r, c-1)] \times [\mathbf{X}_t(r+1, c) - \mathbf{X}_t(r-1, c)]}{\|[\mathbf{X}_t(r, c+1) - \mathbf{X}_t(r, c-1)] \times [\mathbf{X}_t(r+1, c) - \mathbf{X}_t(r-1, c)]\|_2}.$$

A point is valid if it satisfies the human-mask condition (default value 0), has confidence ≥ 0.2, and has finite coordinates. We flip the normal when $\mathbf{n}_{t,r,c}^\top \mathbf{X}_t(r, c) > 0$, orienting it toward the camera. Normals are computed independently per frame, kept aligned with the corresponding XYZ and RGB samples.

Correspondence selection and scale direction. Lifted correspondences are formed only from accepted static pixels with valid 3D predictions. For each cross-window observation, candidate point pairs with degenerate distance denominators are excluded. If two windows represent the same static geometry, their common-scale pairwise distances satisfy

$$a_i\|\mathbf{p}_m - \mathbf{p}_n\|_2 \simeq a_j\|\mathbf{q}_m - \mathbf{q}_n\|_2. \tag{21}$$

Consequently, the ratio in Eq. 8 estimates a_j/a_i, yielding the residual $\log(a_j/a_i) - \log r_{ij}$ in Eq. 9.

Robust observations and reliability. Scale dispersion is measured in the log domain within each matched frame pair and across frame-pair estimates at a window boundary. We use MAD-based dispersion, where median absolute deviation (MAD) measures the median absolute distance from the sample median. The scale-edge weight combines a source-dependent base weight b_{source} with geometric reliability c_{ij}:

$$w_{ij} = \max\big(b_{\mathrm{source}}c_{ij}, 10^{-12}\big). \tag{22}$$

For adjacent-window RoMaV2 edges, reliability is the geometric mean of four dimensionless quality scores, each clipped to $[10^{-4}, 1]$:

$$\begin{aligned} c_{ij} &= \left[\prod_{h=1}^{4} \mathrm{clip}(q_h, 10^{-4}, 1)\right]^{1/4}, \\ q_1 &= \min\left(\frac{K_{\mathrm{kept}}}{K_{\max}}, 1\right), & q_2 &= \frac{1}{1 + (\sigma_{\mathrm{boundary}}/0.03)^2}, \\ q_3 &= \min\left(\frac{\mathrm{median}_p\, v_p}{0.75}, 1\right), & q_4 &= \frac{1}{1 + (\mathrm{median}_p\, \sigma_p/0.05)^2}. \end{aligned} \tag{23}$$

Here K_{kept} counts frame pairs retained after boundary-scale filtering, and $K_{\max}$ is the frame-pair count used for normalization. For frame pair p, v_p is the smaller spatial coverage of valid matches in its two images, and σ_p is the MAD-based dispersion of log 3D distance ratios within that pair. The quantity $\sigma_{\mathrm{boundary}}$ measures dispersion across retained frame-pair log-scale estimates. The medians

in q_3 and q_4 are taken over those retained frame pairs; the numerical constants are implementation defaults.

The default base weights are 1.0 for adjacent RoMaV2 edges, 0.08 for overlap fallback edges, and 0.50 for loop edges. At an adjacent-window boundary, an accepted RoMaV2 scale observation takes precedence; overlap supplies a scale observation only when RoMaV2 fails. Overlap reliability uses valid-point support, spatial coverage, and log-scale dispersion. Loop scale observations for the same window pair are aggregated into one scale edge. Their reliability uses disagreement between the two scale estimates, relative residual, and inlier ratio; aggregation additionally accounts for observation dispersion and retained count. An unsupported scale relation contributes no scale edge, even when an overlap pose factor connects the same windows.

Robust regularization and scale reference. All scale residuals, including the smoothness and prior residuals, are passed together to SciPy's `least_squares` with `loss="soft_l1"`. Using a squared-residual argument $z \geq 0$, we write the corresponding per-residual objective as

$$\rho_s(z) = \eta^2 \left(\sqrt{1 + \frac{z}{\eta^2}} - 1 \right), \tag{24}$$

where z is a weighted squared residual and $\eta > 0$ is the robust scale. For N_w windows indexed from zero, we fix $a_0 = 1$ and define the scale regularizers as

$$\begin{aligned} \mathcal{L}_\mathrm{smooth} &= \sum_{i=1}^{N_\mathrm{w}-2} \rho_s\big(\lambda_s(\log a_{i+1} - 2\log a_i + \log a_{i-1})^2\big), \\ \mathcal{L}_\mathrm{prior} &= \sum_{i=1}^{N_\mathrm{w}-1} \rho_s\big(\lambda_p(\log a_i)^2\big), \qquad a_0 = 1. \end{aligned} \tag{25}$$

Smoothness acts on consecutive window triplets without normalization by frame spacing, encouraging neighboring scale ratios a_{i+1}/a_i and a_i/a_{i-1} to agree. The prior favors unit correction $a_i = 1$. Both regularizer weights enter the squared residual before robustification, matching the data term in Eq. 9. We use $\lambda_s = 0.03$, $\lambda_p = 0.002$, and $\eta = 0.03$.

Fixing $a_0 = 1$ sets the scale reference. A component disconnected in the observation graph has no measured scale relation to that reference, even if regularization provides an estimate.

Rigid factors after scale synchronization. Let $\mathbf{G}_k = [\mathbf{R}_k^G, \mathbf{d}_k^G]$ denote the rigid transform from the scale-normalized window frame to the global reconstruction frame, with the homogeneous bottom row implicit. A local point $\mathbf{p}$ and a local camera pose $[\mathbf{R}_{k,t}, \mathbf{t}_{k,t}]$ transform as follows, with all positions still in reconstruction units:

$$\begin{aligned} \mathbf{p}^w &= \mathbf{R}_k^G(a_k\mathbf{p}) + \mathbf{d}_k^G, \\ \mathbf{R}_t^{(k)} &= \mathbf{R}_k^G\mathbf{R}_{k,t}, \\ \mathbf{t}_t^{(k)} &= \mathbf{R}_k^G(a_k\mathbf{t}_{k,t}) + \mathbf{d}_k^G. \end{aligned} \tag{26}$$

For a pose factor e connecting windows i and j, $\mathbf{Z}_e$ maps the scale-normalized frame of j to that of i, consistent with $\mathbf{G}_i^{-1}\mathbf{G}_j$. One reference pose fixes the rigid gauge; the index e distinguishes factors sharing the same window pair.

Pose-factor construction and multiplicity. Adjacent windows are connected by an overlap factor and, when geometric verification succeeds, an additional 3D-alignment factor. The latter comes from one fixed-scale rigid fit to pooled matches across the retained RoMaV2 frame pairs. Verified loop observations remain separate pose factors, including multiple observations for one window pair. Default loop retrieval excludes the immediately preceding window and therefore connects non-adjacent windows.

Frame-pair planning avoids duplicate RoMaV2 frame pairs, but factor insertion does not deduplicate shared matches across sources or model their correlation. Thus, loop observations are aggregated per pair in the scale graph but remain separate in the pose graph.

Scale and pose verification are separate acceptance decisions; SALAD retrieval alone contributes neither type of constraint.

A.4 World-Space Translation Refinement

Optimization variables and reprojection. Let $\mathbf{J}^w_{t,j}$ and $\mathbf{X}^w_{t,v}$ denote SMPL-X body joints and vertices before refinement, expressed in body-model units. These world-space body joints are distinct from the root-relative skeletal reference $\mathbf{J}_t$ used for scale reasoning. All joints and vertices of a frame receive the same residual translation:

$$\begin{aligned} \mathbf{J}^{w,\mathrm{ref}}_{t,j} &= \mathbf{J}^w_{t,j} + \boldsymbol{\delta}^w_t, \\ \mathbf{X}^{w,\mathrm{ref}}_{t,v} &= \mathbf{X}^w_{t,v} + \boldsymbol{\delta}^w_t. \end{aligned} \tag{27}$$

For perspective projection $\pi_{\mathbf{K}_t}$, the refined joint projects as

$$\mathbf{u}^{\mathrm{ref}}_{t,j} = \pi_{\mathbf{K}_t}\left[\mathbf{R}_t^\top\left(\mathbf{J}^{w,\mathrm{ref}}_{t,j} - s_g\mathbf{t}_t\right)\right]. \tag{28}$$

The subtraction uses the metric camera translation $s_g\mathbf{t}_t$; the body joints are already metric.

Objective terms and contact gating. The terms in Eq. 12 compare reprojections with image observations, regularize temporal motion, enforce accepted support contacts, and penalize deviations from the initial translation. Predicted contact is activated only after its geometric consistency with a static world is checked. Only $\{\boldsymbol{\delta}^w_t\}$ are optimized; body parameters, cameras, scene geometry, and scales remain fixed, preserving all within-body distances.

B Evaluation Details

Benchmarks and reporting. Human reconstruction and ablations use EMDB2, with additional human comparisons on RICH. Camera evaluation covers 25 EMDB2 sequences and the eight Bonn RGB-D sequences listed in Table 8, using monocular RGB input. For the camera-space human evaluation on 3DPW, we follow the 14-joint setup and evaluation protocol of OnlineHMR (Zhao et al., 2026). We follow the official 3DPW train/test split for the evaluation in Table 7. All test sequences were strictly excluded from video-head training. The two-epoch backbone and scale pilots are separate experiment groups. Percentage improvements are computed from the displayed values.

Published reference metric conventions. The public GVHMR evaluator uses a rigid fit with scale fixed to one for RTE, then normalizes root-position errors by the ground-truth trajectory's total path length and reports percentages (Shen et al., 2024). For W-MPJPE$_{100}$ and WA-MPJPE$_{100}$, it fits a similarity transform using the first two frames and the entire segment, respectively. Consequently, the component of an error removed by similarity alignment can still contribute to RTE. These alignment conventions must be distinguished from the method's own metric-scale prediction. The evaluator's foot-skating test uses a 10 mm inter-frame displacement threshold on ground-truth foot vertices; this evaluation threshold is distinct from HER's predicted-contact and geometric gates during optimization.

Sources of comparison values. Baseline entries are published results. For EMDB2, the WHAM, GVHMR, TRAM, PromptHMR-Vid, JOSH, Human3R, JOSH3R, and OnlineHMR motion values are tabulated in Table 2 of OnlineHMR (Zhao et al., 2026); its WHAM and GVHMR entries use DPVO, which the † symbol denotes for EMDB2 in Table 1. Human3R's global-motion table also reports the RICH motion values for GVHMR, TRAM, JOSH, and Human3R (Chen et al., 2026b). The RICH WHAM motion values and both SHOW motion results are reported in Table 2 of SHOW (Shi et al., 2026). The UniSH motion values appear in its Table 2 (Li et al., 2026); UniCon3R's Table 1 supplies its own results and the RICH JOSH3R values (Sur et al., 2026). For Bonn, the DROID-SLAM, MonST3R, MegaSaM, and Pi3MOS-SLAM entries appear in Table 1 of Pi3MOS-SLAM (Zhong et al., 2025).

C Additional Quantitative Results and Analysis

C.1 Camera-Space Human Reconstruction on 3DPW

Table 7 compares camera-space human reconstruction on the official 3DPW test split, which is excluded from video-head training. HER has the lowest reported PA-MPJPE, MPJPE, and PVE

among the listed methods. These local reconstruction metrics complement the world-space motion results in Table 1; they do not measure camera-trajectory accuracy or scene-scale recovery.

Table 7: **Camera-space human reconstruction on 3DPW.** Bold and underlined values mark the best and second-best results.

Method	PA-MPJPE ↓	MPJPE ↓	PVE ↓
Human3R (Chen et al., 2026b)	44.1	71.2	84.9
UniSH (Li et al., 2026)	48.8	75.6	88.8
OnlineHMR (Zhao et al., 2026)	<u>43.7</u>	<u>69.9</u>	<u>83.7</u>
HER (Ours)	**36.0**	**55.4**	**67.8**

C.2 Bonn RGB-D Benchmark

Table 8 reports per-sequence camera ATE on Bonn RGB-D (Palazzolo et al., 2019), using monocular RGB input.

Table 8: **ATE RMSE (cm) ↓ evaluation on Bonn RGB-D.** All methods are monocular. Red and orange mark the best and second-best value in each column.

Method	w/o Calib.	Balloon	Balloon2	Crowd	Crowd2	Person	Person2	Moving	Moving2	Avg.
DROID-SLAM (Teed & Deng, 2021)	✗	7.50	4.10	5.20	6.50	4.30	5.40	2.30	4.00	4.91
TTT3R (Chen et al., 2026a)	✓	21.50	15.40	9.80	7.70	30.00	21.40	33.40	41.20	22.55
MonST3R (Zhang et al., 2025)	✓	5.40	7.20	5.40	6.90	11.90	11.10	3.30	7.40	7.30
VGGT-SLAM 1.0† (Maggio et al., 2025)	✓	5.39	6.31	4.11	4.91	9.37	11.05	7.75	8.31	7.15
VGGT-SLAM 2.0† (Maggio & Carlone, 2026)	✓	3.58	3.91	2.06	3.57	6.33	4.29	4.79	3.92	4.05
MegaSaM (Li et al., 2025b)	✓	3.70	2.60	1.60	7.20	4.10	4.00	1.40	3.40	3.51
Pi3MOS-SLAM (Zhong et al., 2025)	✗	2.60	2.20	1.30	2.20	3.20	3.00	1.20	1.90	2.20
Ours	✓	1.74	1.68	1.28	1.96	2.96	5.39	1.42	2.11	2.32

Calib.: supplied camera calibration; †full-frame configuration. Person and Moving denote the person-tracking and moving-box sequences; the suffix 2 denotes the second sequence of each family. Values are rounded to two decimals; Avg. retains each reported aggregate.

Per-sequence behavior. HER achieves the lowest ATE on Balloon, Balloon2, Crowd, Crowd2, and Person. Its average of 2.32 cm is the lowest among the listed methods without supplied calibration and second overall to Pi3MOS-SLAM at 2.20 cm. The largest deficit relative to Pi3MOS-SLAM occurs on Person2 (5.39 versus 3.00 cm), despite HER's lower errors on five sequences. Thus, the number of per-sequence wins does not imply the lowest mean error across the benchmark.

C.3 Scale Sensitivity and the GT-Scale Oracle

The oracle obtains scale by aligning the estimated camera poses with ground-truth camera poses and substitutes that alignment scale into Points Align. Thus, scale is calibrated using camera ground truth rather than fitted to the human trajectory. This provides an oracle metric calibration for the Points Align baseline; it does not replace HER's learned localization with ground-truth human positions. Replacing only the scale of Points Align reduces RTE from 1.18 to 0.57, a 51.7% reduction. The oracle is also 50.4% below full HER's 1.15, but that comparison changes both localization and scale; the scale-only comparison is against Points Align. Scale affects the magnitude of world-space displacement even when articulation is accurate.

Scale sensitivity of world-space composition. To interpret the role of the two scales in HER, define its unscaled camera-space translation as $\boldsymbol{\tau}_t^{\mathrm{raw}} = \boldsymbol{\mu}_t + g_\tau(\mathbf{q}_t^{\mathrm{trans}\prime})$. Before final refinement, Eq. 19 gives the world-space pelvis position

$$\mathbf{r}_t^w = \mathbf{R}_t\big(\mathbf{p}_\beta + s_t^{\mathrm{e}}\boldsymbol{\tau}_t^{\mathrm{raw}}\big) + s_g\mathbf{t}_t. \tag{29}$$

Holding camera poses, body parameters, and $\boldsymbol{\tau}_t^{\mathrm{raw}}$ fixed, a perturbation of the two scales changes this position by

$$\Delta\mathbf{r}_t^w = \mathbf{R}_t\boldsymbol{\tau}_t^{\mathrm{raw}}\,\Delta s_t^{\mathrm{e}} + \mathbf{t}_t\,\Delta s_g. \tag{30}$$

This is an exact finite-difference identity for the initial composition, not a model of the response after translation refinement. The first term changes camera-relative human placement; the second changes the camera contribution to world placement. When s_t^{e} varies independently of s_g, their effect is generally not a single similarity transform of the reconstructed human motion, because metric body dimensions remain fixed.

What the oracle establishes. Under the published reference conventions in Appendix B, similarity alignment can remove a component of global scale error, whereas the rigid alignment used for RTE cannot rescale the trajectory. This distinction is consistent with the oracle's larger change in RTE: W-MPJPE$_{100}$ changes only from 173.5 to 173.1 and WA-MPJPE$_{100}$ from 65.7 to 64.6; jitter and foot skating are unchanged at the displayed precision. The controlled comparison identifies scale sensitivity in Points Align. It does not quantify the fraction of HER's error caused by scale, or establish uniformly better motion quality: HER retains lower joint errors. The oracle uses test-time camera ground truth and is an analysis reference rather than a deployable variant.

C.4 Temporal Reasoning and Translation Refinement

Table 3 provides two distinct comparisons. With temporal reasoning enabled, adding final translation refinement reduces W-MPJPE$_{100}$ from 167.4 to 161.1 and WA-MPJPE$_{100}$ from 66.1 to 61.3. Jitter decreases from 129.9 to 12.6 (90.3%), and foot skating from 17.1 to 5.3 (69.0%); RTE changes from 1.16 to 1.15. These results show the contribution of the complete refinement objective in Eq. 12; they do not isolate its individual reprojection, temporal, contact, or initialization terms.

With final refinement enabled, adding temporal reasoning reduces W-MPJPE$_{100}$ from 163.8 to 161.1 and WA-MPJPE$_{100}$ from 62.6 to 61.3. Jitter changes from 12.7 to 12.6 and foot skating from 5.4 to 5.3. Thus, this comparison measures the benefit of temporal reasoning within the refined pipeline.

The skeletal-reference and video-head comparisons also require distinct controls. Comparing HER with SMPL-X Joints changes the skeletal reference while retaining the retrained head. Comparing SMPL-X Joints with PHMR-Vid Head retains the SMPL-X-regressed reference and changes the video-head configuration. Comparing PHMR-Vid Head directly with HER changes both factors.

C.5 Point-Cloud Backbone Configuration and Parameter Counts

Encoder size. Table 9 reports encoder counts only, separately from accuracy configurations. The backbone and scale-reasoning pilots are different experiment groups; their HER rows do not form a single ablation chain.

Table 9: **Encoder parameter counts before and after trimming.** Reduction is computed from the displayed counts. The mapping of trimmed checkpoints to Table 10 reports encoder parameter counts before and after trimming, together with the resulting percentage reductions.

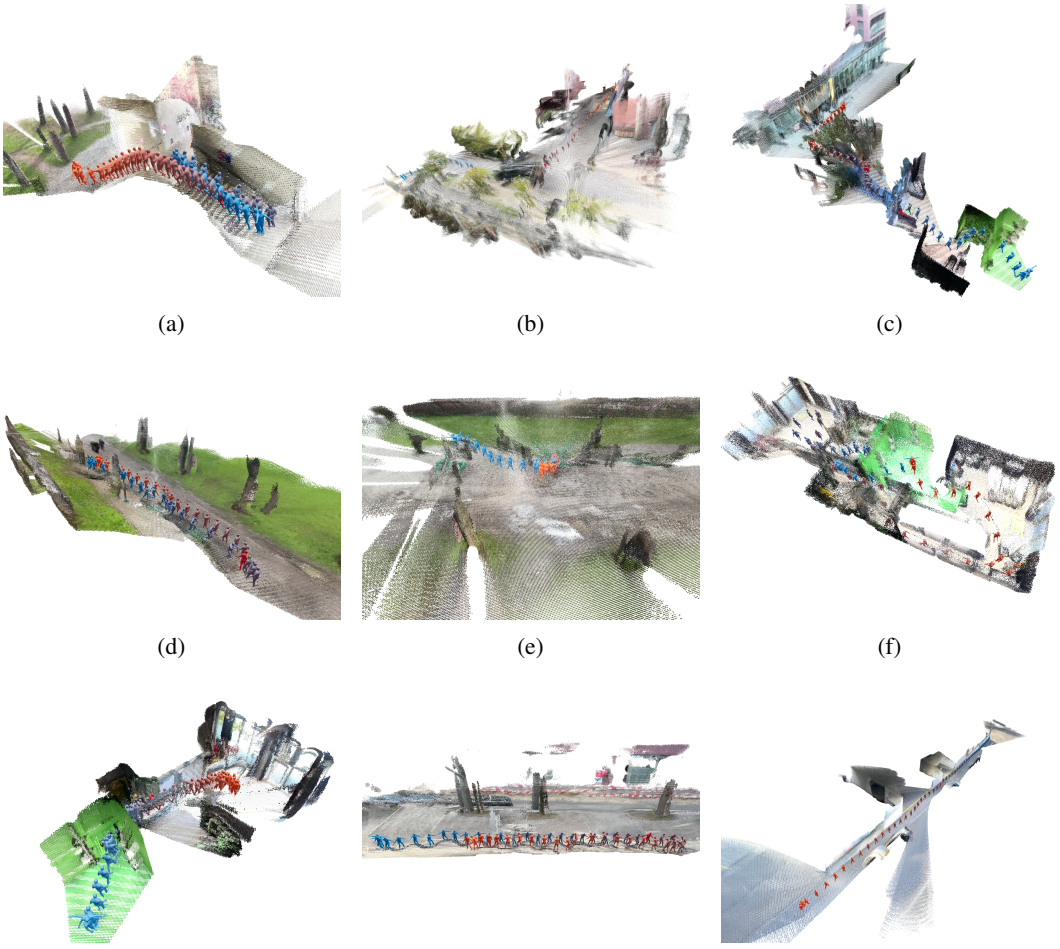
Model	Encoder params	Params after trimming	Reduction
PTv3	73.13M	41.43M	43.35%
LitePT-S	12.44M	6.20M	50.16%
LitePT-B	44.44M	32.22M	27.50%
LitePT-L	84.72M	57.30M	32.37%
Utonia	137.25M	97.16M	29.21%

Table 10 consolidates the two-epoch EMDB2 backbone studies without final translation refinement. A stage is a resolution level containing the listed attention blocks. Four-stage Utonia has the lowest joint errors, with RTE tied with three-stage Utonia at the displayed precision. Trimming PTv3 or LitePT increases joint errors in these runs. Fine-tuning four-stage Utonia improves WA-MPJPE$_{100}$ from 83.0 to 71.5 and W-MPJPE$_{100}$ from 188.2 to 177.3 relative to freezing it. These observations support the selected configuration under the pilot budget. Because these experiments use a fixed two-epoch pilot budget, we interpret them as configuration studies rather than converged architecture comparisons.

Table 10: **Point-cloud backbone, depth, and fine-tuning comparison.** Two-epoch EMDB2 results without post-processing. All backbones are trainable except the frozen row. Bold marks the lowest displayed error, including ties.

Backbone	Stages	Blocks per stage	W-MPJPE$_{100}$ ↓	WA-MPJPE$_{100}$ ↓	RTE ↓
PTv3 (full)	5	[3, 3, 3, 6, 3]	189.9	83.9	1.31
PTv3 (drop)	4	[3, 3, 3, 6]	190.3	84.7	1.31
LitePT-S (full)	5	[2, 2, 2, 6, 2]	180.0	75.2	1.29
LitePT-S (drop)	4	[2, 2, 2, 6]	181.7	76.0	1.29
LitePT-B (full)	5	[3, 3, 3, 12, 3]	179.6	75.5	1.30
LitePT-B (drop)	4	[3, 3, 3, 12]	182.5	77.6	1.30
LitePT-L (full)	5	[3, 3, 3, 12, 3]	179.7	74.9	1.29
LitePT-L (drop)	4	[3, 3, 3, 12]	181.1	76.4	1.30
Utonia (full)	5	[3, 3, 3, 12, 3]	177.4	72.6	1.29
Utonia (drop 2 stages)	3	[3, 3, 3]	177.9	72.5	**1.28**
Utonia (frozen, drop 1)	4	[3, 3, 3, 12]	188.2	83.0	1.30
Utonia (drop 1, ours)	4	[3, 3, 3, 12]	**177.3**	**71.5**	**1.28**

D ADDITIONAL QUALITATIVE RESULTS ON EMDB2



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5: **Qualitative human–environment reconstruction on EMDB2.** Complete sequence-level reconstructions produced by HER in a shared metric world coordinate system.

Figure 5 presents nine qualitative human–environment reconstructions on EMDB2. Each visualization shows the reconstructed human motion together with the recovered environment in the shared metric world coordinate system. Human trajectories are color-coded over time, progressing from blue to red.

Comparison with other methods. Figure 6 compares HER with Human3R (Chen et al., 2026b) and OnlineHMR (Zhao et al., 2026). HER more closely follows the ground-truth trajectories, particularly in the second and third examples.

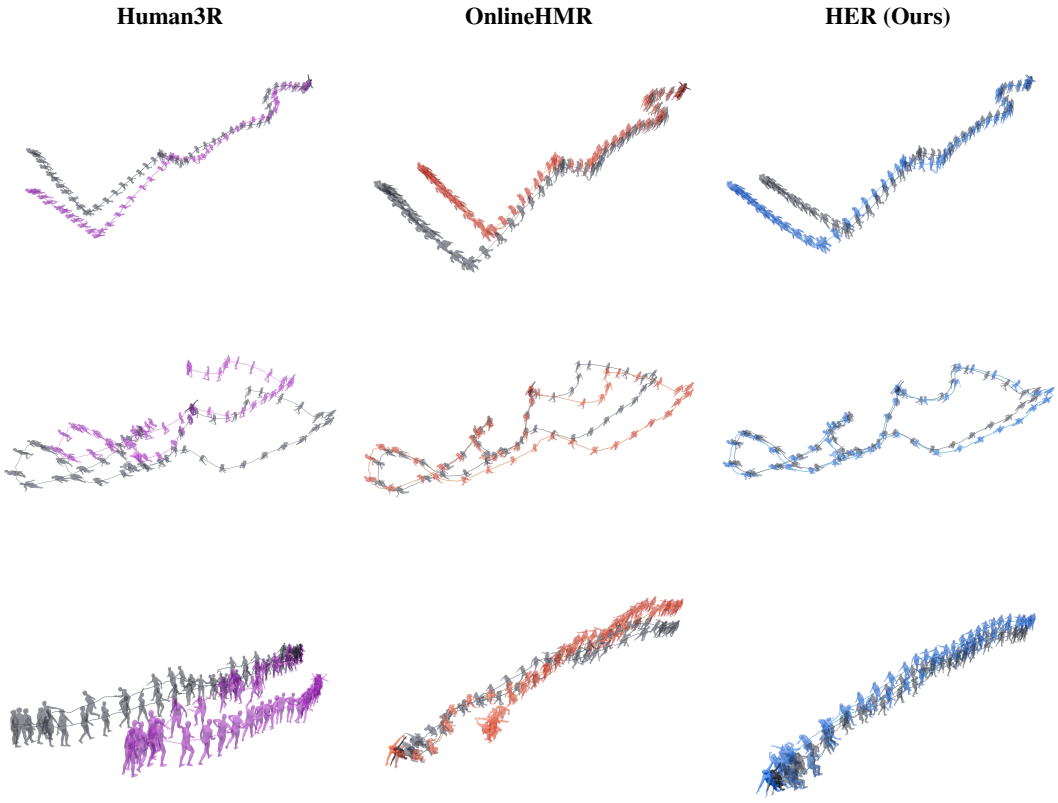


Figure 6: **Comparison with other methods on EMDB2.** Each row shows one example. Columns show Human3R (purple), OnlineHMR (orange), and HER (blue), with ground truth in gray. For each method, a rigid transform is fitted to corresponding predicted and ground-truth 3D human joints in the first frame and applied to the entire sequence, with scale fixed to one.

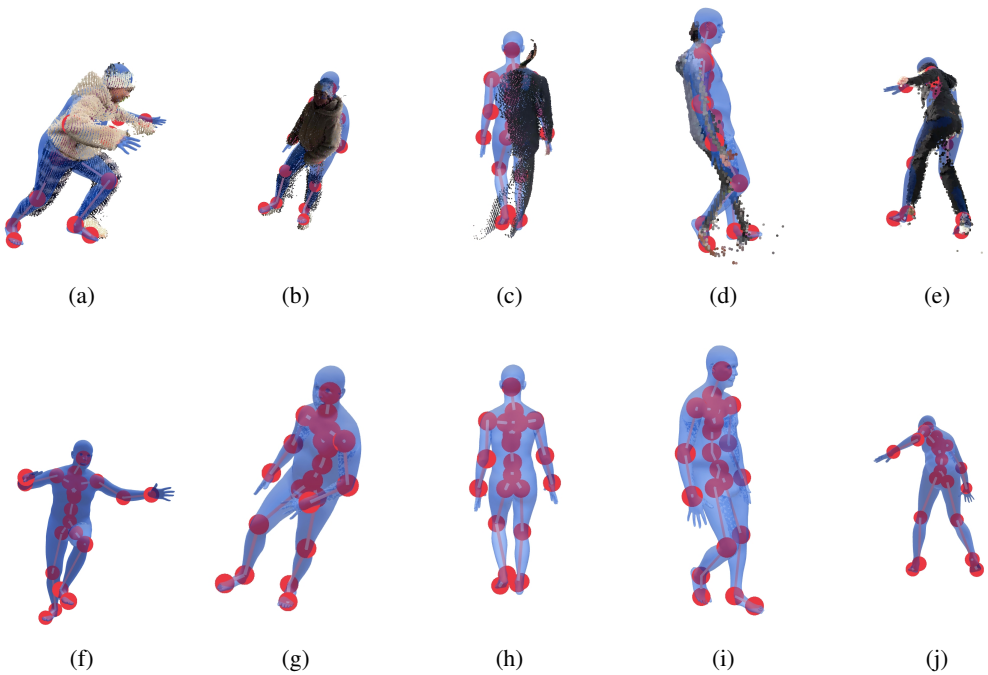


Figure 7: **Close-up reconstruction on an EMDB2 sequence.** Top row (a-e): Reconstructed human point clouds together with the skeletal references predicted by the video head. Bottom row (f-j): Final predicted human meshes after the complete HER pipeline.