HARMONI: ALIGNING HUMAN AND SCENE PRIORS FOR MULTI-VIEW 4D RECONSTRUCTION

Anonymous authors
Paper under double-blind review



(a) Reconstructed humans and scene (b) Comparison

Figure 1: (a) We present **HARMONI**, which reconstructs cameras, scene, and humans with identities from multi-view or monocular videos without test-time optimization. (b) With *bidirectional anchoring*, our model outperforms previous human-scene reconstruction methods in both settings.

ABSTRACT

Recent advances in 3D foundation models have enabled joint reconstruction of humans and their surrounding environments. However, combining independently trained human and scene priors often produces misalignment in scale and depth. We observe that the two priors have complementary strengths. The scene prior provides consistent depth but approximate scale, while the human prior provides fixed body scale but less reliable depth. Motivated by this observation, we present **HARMONI**, a feed-forward framework that reconstructs cameras, scene, and humans with identities from monocular or multi-view video without test-time optimization. We introduce *bidirectional anchoring*, in which scene depth guides human placement, while human keypoints calibrate the scene scale to match the human body. While most existing approaches target monocular inputs and multi-view methods rely on optimization or re-identification, our approach naturally extends to multiple views. Building on bidirectional anchoring, we introduce a *multi-view fusion module* that merges per-view estimates, while reducing the influence of unreliable views and visually similar individuals. Experiments show that our framework outperforms previous human-scene methods in global motion and multi-view pose estimation by up to 28% and 64%, while running 28× faster than optimization-based approaches.

1 INTRODUCTION

Understanding how people interact with 3D environments requires reconstructing both humans and their surroundings in a shared coordinate system. Joint reconstruction of humans and scene enables

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.

downstream applications such as training humanoids (Allshire et al., 2025; Wang et al., 2026c), simulating scenes for autonomous driving (Chen et al., 2025), and creating AR/VR entertainment (Kim et al., 2025; Hwang et al., 2026). With advances in 3D foundation models such as DUSt3R (Wang et al., 2024a) and VGGT (Wang et al., 2025a), many works have tried to jointly recover humans and their environments by combining 3D foundation models with human mesh recovery models (Baradel et al., 2024). Recent approaches such as UniSH (Li et al., 2026) and Human3R (Chen et al., 2026) unify human and scene reconstruction into a single feed-forward architecture.

However, one of the challenges in jointly reconstructing humans and scene is that the two can be misaligned. This issue arises in two aspects: scale and depth. Since 3D foundation models *estimate* the metric scale of the scene, it can differ from the scale of parametric human meshes such as SMPL (Loper et al., 2015), which are *defined* in metric units. In addition, because the scene and human priors are trained independently, the depth of the scene points corresponding to the human can differ from the depth given by the z-component of the SMPL global translation. If the scale is inconsistent, reconstructed humans can float above or penetrate the ground, and if the depth is inconsistent, they can be placed too close to or too far from the camera relative to the scene (Fig. 1b).

Moreover, most existing works target monocular inputs, and extending to multi-view introduces additional challenges. Predictions from different views must be expressed in a shared coordinate system, with humans and scene aligned within each view and person correspondences established across views. Existing multi-view approaches address these requirements through additional stages. HSfM (Müller et al., 2025) relies on external 2D keypoint estimation, known cross-view person identities, and iterative optimization, whereas HAMSt3R (Rojas et al., 2025) obtains SMPL meshes through a fitting procedure. These stages increase system complexity and computational cost.

To address these issues, we present **HARMONI** (aligning **H**uman **A**nd scene p**R**iors for **M**ulti-view 4D Rec**ON**struct**I**on), a feed-forward framework that reconstructs cameras, the scene, and humans with identities from monocular or multi-view videos without test-time optimization (Fig. 1a). We observe that human and scene priors have complementary strengths. The scene prior provides consistent depth but only approximate scale, whereas the human prior provides a fixed body scale but less reliable depth. Motivated by this observation, we introduce *bidirectional anchoring* that exploits these complementary strengths. Human keypoints calibrate the scene scale to the human body, while scene depth guides global human placement. This alignment places humans coherently within the scene and improves both global motion and multi-view pose accuracy (Fig. 1b).

Since bidirectional anchoring aligns human and scene predictions within each view, our framework can be naturally extended to multi-view inputs. We introduce a *multi-view fusion module* that aggregates per-view human estimates into a global representation. Specifically, confidence-weighted averaging suppresses unreliable per-view estimates, triangulation recovers precise global positions, and a refine transformer refines the fused estimates. Our module establishes cross-view person correspondences through geometry-based association using 3D positions and human poses. Unlike appearance-based re-identification (Li et al., 2024), our association remains robust for visually similar individuals. Together, our module enables optimization-free multi-view multi-person fusion.

Experiments on a monocular (Kaufmann et al., 2023) dataset and multi-view (Huang et al., 2022; Khirodkar et al., 2023; Grauman et al., 2024) datasets demonstrate that our framework outperforms previous human–scene reconstruction methods, reducing global motion error by up to 28% and multi-view pose error by up to 64%, while running 28$\times$ faster than optimization-based baselines.

Our main contributions are summarized as follows:

- We present **HARMONI**, a feed-forward framework that recovers cameras, scenes, and humans with identities from monocular and multi-view video without test-time optimization.
- We propose *bidirectional anchoring*, which aligns human and scene priors: human keypoints calibrate scene scale, while scene depth guides human placement.
- We introduce a *multi-view fusion module* that merges per-view estimates without test-time optimization, while reducing the influence of unreliable views.
- Our framework reduces global motion and multi-view pose errors by up to 28% and 64%, respectively, and runs over 28$\times$ faster than optimization-based approaches.

2 Related Work

3D Scene Reconstruction. Classical 3D reconstruction approaches such as Structure-from-Motion (SfM) (Longuet-Higgins, 1981; Tomasi & Kanade, 1992; Beardsley et al., 1996; Fitzgibbon & Zisserman, 1998; Schonberger & Frahm, 2016) and Multi-View Stereo (MVS) (Seitz et al., 2006; Furukawa & Ponce, 2009; Schönberger et al., 2016) estimate camera poses and scene geometry through feature matching and bundle adjustment. While highly accurate, these optimization-based pipelines are computationally expensive and struggle in dynamic scenes. Recent data-driven approaches replace iterative optimization with feed-forward neural architectures. DUSt3R (Wang et al., 2024a) predicts a 3D point map from an image pair in the first camera's coordinate system, enabling efficient reconstruction, and MASt3R (Leroy et al., 2024) improves matching stability and reconstruction accuracy. VGGT (Wang et al., 2025a) replaces explicit camera estimation with direct network prediction, while Pi3 (Wang et al., 2026b) introduces a permutation-equivariant architecture to improve robustness. MapAnything (Keetha et al., 2026) and Depth Anything 3 (Lin et al., 2026) extend feed-forward reconstruction toward metric-scale prediction. More recently, VGGT-Ω (Wang et al., 2026a) further improves reconstruction accuracy and efficiency for both static and dynamic scenes.

Human Mesh Recovery. Human Mesh Recovery (HMR) predicts parameters of body models such as SMPL (Loper et al., 2015) and SMPL-X (Pavlakos et al., 2019) from images. HMR2.0 (Goel et al., 2023) employs a fully transformer-based architecture for single-image HMR, while Multi-HMR (Baradel et al., 2024) enables single-shot multi-person whole-body reconstruction. Multi-HMR 2 (Fiche et al., 2026) further adopts a DETR-style decoder with camera intrinsics prediction. While these methods operate in the camera coordinate frame, recent research has shifted toward reconstructing humans in the world coordinate system. SLAHMR (Ye et al., 2023) separately reconstructs camera and human motion and then aligns them in a shared world coordinate system through optimization. WHAM (Shin et al., 2024) improves world-grounded motion estimation through motion priors and contact-aware refinement. GVHMR (Shen et al., 2024) introduces a gravity-aligned intermediate coordinate system to stabilize world-space motion prediction. TRAM (Wang et al., 2024b) combines SLAM-based camera estimation with transformer-based motion regression to reconstruct metric-scale human trajectories in world coordinates. These approaches enable world-grounded human motion reconstruction from monocular video. However, they focus on human motion rather than jointly reconstructing humans and surrounding scene geometry.

Human-Scene Reconstruction. Recent work attempts to combine 3D foundation models with human mesh recovery to jointly reconstruct humans and their surroundings. JOSH (Liu et al., 2026b) reconstructs humans and scenes from monocular videos using foot-contact cues for global alignment, and its extension JOSH3R further adopts a feed-forward architecture. UniSH (Li et al., 2026) integrates Pi3 and CameraHMR (Patel & Black, 2025), and finetunes the backbone and trains AlignNet, which requires large-scale internet data. Human3R (Chen et al., 2026) targets unified human-scene reconstruction, but directly regresses human translation and relies on the scale estimate of CUT3R (Wang et al., 2025b). MetricHMSR (Song et al., 2026) and GRAFT (Pradyumna et al., 2026) recover humans and scenes from a single image, but do not model cameras. Since these approaches operate in monocular settings, several works explore multi-view reconstruction. HSfM (Müller et al., 2025) uses human joints as cross-view correspondences to recover cameras and metric scale, but does so through iterative optimization with external 2D keypoints and requires known person identities. HAMSt3R (Rojas et al., 2025) predicts segmentation, dense pose, and scene geometry in a feed-forward manner at a single timestep, but requires additional fitting to obtain SMPL meshes. Trophies (Liu et al., 2026a) handles multi-view video, but still relies on test-time optimization for cross-view alignment. While HSfM and JOSH exploit humans for global alignment, they rely on test-time optimization. In contrast, our bidirectional anchoring calibrates the scene scale from body proportions and anchors human depth on the scene prior in a single feed-forward pass, providing aligned per-view estimates for our optimization-free multi-view fusion.

3 Method

We present **HARMONI**, a feed-forward framework that reconstructs cameras, scene, and humans with identities from monocular or multi-view videos without test-time optimization. Given RGB

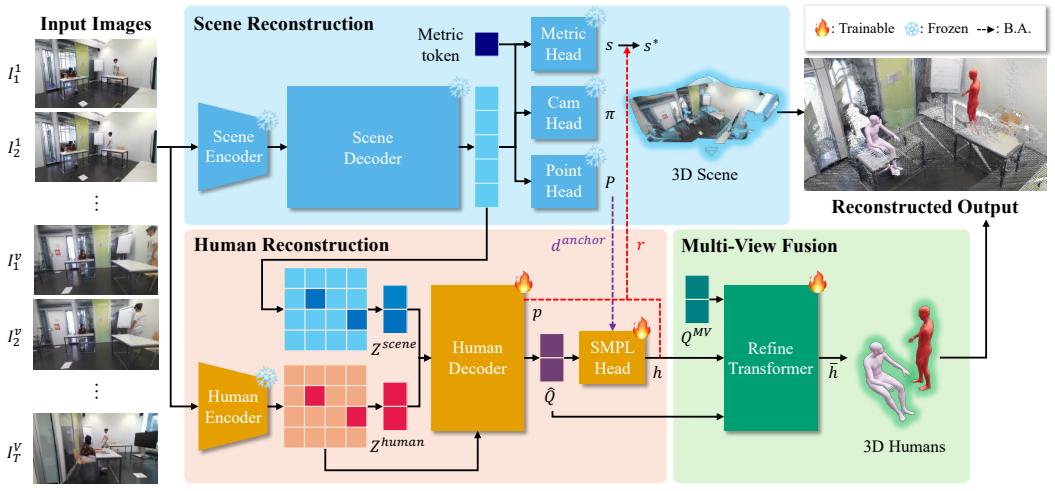

Figure 2: **Overview of our framework.** The scene encoder and decoder reconstruct cameras π, point maps P, and an approximate metric scale s. The human decoder and SMPL head predict per-view human estimates h. Bidirectional anchoring (B.A.) couples the two branches. Scene depth d^{anchor} guides human placement, and the keypoint-based ratio r calibrates the scene scale. For multi-view inputs, the multi-view fusion module merges per-view estimates into a global estimate $\bar{h}$.

frames $\{I_t^v\}_{t=1..T}^{v=1..V}$ from V views and T timesteps, the model estimates per-frame camera-space point maps P_t^v and camera parameters π_t^v, as well as the SMPL-X parameters of M people in a shared world coordinate system. For person m at time t, the global human representation is $\bar{h}_{t,m} = (\bar{\theta}_{t,m}, \bar{\beta}_m, \bar{R}_{t,m}, \bar{\tau}_{t,m})$, where $\bar{\theta}_{t,m} \in \mathbb{R}^{52\times3}$ denotes canonical pose parameters, $\bar{\beta}_m \in \mathbb{R}^{10}$ shape parameters, $\bar{R}_{t,m} \in SO(3)$ global root rotation, and $\bar{\tau}_{t,m} \in \mathbb{R}^3$ global head translation. Bars denote final estimates in the world coordinate system. Fig. 2 shows the pipeline overview.

3.1 Model Architecture

HARMONI combines a scene prior and a human prior within a single architecture. For human modeling, we adopt SMPL-X (Pavlakos et al., 2019), a parametric body model controlled by the pose parameters θ and the shape parameters β.

Scene Reconstruction. We use frozen Pi3X (Wang et al., 2026b) for scene reconstruction. We observe that human reconstruction can benefit from the geometric accuracy of the 3D foundation model, and fine-tuning the scene prior on human datasets can degrade scene quality. We flatten the input frames into a sequence $\{I_n\}_{n=1..N}$, where $N = T \times V$ and process them with the scene encoder and decoder. The decoder jointly processes all frames with alternating attention and feeds its tokens to camera and point heads that predict π_n and P_n. A metric token additionally predicts a global scale s, which is applied to the point maps and camera translations to obtain approximate metric-scale geometry. The decoded tokens are reshaped into feature maps $\{\hat{F}_n^{\text{scene}}\}_{n=1}^N$ and passed to the human reconstruction branch.

Human Reconstruction. For human reconstruction, we adopt the frozen Multi-HMR (Baradel et al., 2024) encoder for human-centric features, while the subsequent human decoder and SMPL head are trained from scratch. We note that the scene encoder alone does not capture fine-grained human details, and merging such heterogeneous representations into a single encoder requires dedicated distillation. In parallel, the human encoder extracts human features F_n^{human} from each frame. Following Baradel et al. (2024), a detection MLP identifies head patches, from which we extract a human token $Z_{n,m}^{\text{human}}$ and the spatially corresponding scene token $Z_{n,m}^{\text{scene}}$ for each detected person m. We concatenate these tokens with a positional embedding to form the person query $Q_{n,m} = [Z_{n,m}^{\text{scene}} \mid Z_{n,m}^{\text{human}}]$. The human decoder refines the queries through self-attention among people in the same frame and cross-attention to F_n^{human}, producing $\hat{Q}_{n,m}$. Finally, the SMPL head predicts $(\theta_{n,m}, \beta_{n,m}, R_{n,m})$ from $\hat{Q}_{n,m}$, while the translation $\tau_{n,m}$ is derived from the scene depth map, as described in Sec. 3.2.



(a) Depth residual prediction (b) Keypoint heatmap (c) Scale calibration

Figure 3: **Bidirectional Anchoring.** (a) Human depth is anchored on the head surface depth with a predicted residual to the head joint. (b) Keypoints for calibration are localized using heatmaps. (c) Scene scale is calibrated by comparing SMPL and image body proportions.

3.2 Bidirectional Anchoring

As shown in Fig. 1b, aligning independently trained human and scene priors is challenging. If the predicted human depth is inconsistent, humans are placed too close to or too far from the camera, and if the scale is inconsistent, they float above or penetrate the ground. We observe that the two priors have complementary strengths. The scene prior provides consistent depth but only an approximate metric scale, whereas the human prior provides body geometry defined in metric units. Based on this observation, we introduce *bidirectional anchoring*, which uses each prior to anchor the other.

Depth Residual Prediction. To place each human in the scene, we first estimate the 3D head joint translation $\tau_{n,m} \in \mathbb{R}^3$ in the camera coordinate system. We choose the head instead of the root because it is the most distinctive body part across individuals and is the least frequently occluded. Instead of directly regressing the translation as in Chen et al. (2026), we predict the 2D head keypoint $p^{\text{head}}_{n,m}$ with an offset MLP from the detected head token $Z^{\text{human}}_{n,m}$, and unproject it with the predicted intrinsics $K_n \in \mathbb{R}^{3\times3}$ and the depth $d_{n,m} \in \mathbb{R}$ (Fig. 3a). We anchor this depth on the scene prior by sampling $d^{\text{anchor}}_{n,m}$ at the head keypoint from the depth map D_n, which is the z-coordinate of the point map P_n. As the 3D head joint lies behind the visible head surface, we predict a positive residual $\Delta d_{n,m}$ from $\hat{Q}_{n,m}$ and add it to the anchor depth. The final 3D head translation is defined as

$$d_{n,m} = d^{\text{anchor}}_{n,m} + \Delta d_{n,m}, \qquad \tau_{n,m} = d_{n,m}\, K_n^{-1} \begin{bmatrix} p^{\text{head}}_{n,m} \\ 1 \end{bmatrix}. \tag{1}$$

We observe that anchoring the depth on the scene prior stabilizes training and improves generalization compared to direct regression.

Scale Calibration. With Eq. 1, the SMPL head joints are precisely positioned in the scene. However, the scene prior estimates the scene at an *approximate* metric scale, causing misalignment with metric-scale SMPL meshes. We therefore calibrate the scene scale using the discrepancy between image-space 2D body proportions and those from the projected SMPL-X mesh. We use the 2D head keypoint predicted by the offset MLP as the calibration reference, and select the spine, pelvis, knees, and feet as additional joints. For each non-head joint k, we first obtain an instance-conditioned heatmap by projecting the refined query $\hat{Q}_{n,m}$ with a joint-specific linear layer and taking its dot product with every patch token in the human feature map F^{human}_n (Fig. 3b). Then, we calculate the 2D keypoint $p^k_{n,m}$ by applying soft-argmax to its heatmap. In addition, two linear heads predict a visibility and a confidence logit for each joint, indicating whether it is truncated or occluded, respectively. We use an offset MLP instead of a heatmap for head localization, because it is more robust in crowded scenes as it predicts the head joint within the instance-specific head token boundary. As shown in Fig. 3c, we compute the image and SMPL head-to-joint lengths for each joint k and derive the calibration ratio as

$$\ell^{\text{img}}_{k,n,m} = \|p^{\text{head}}_{n,m} - p^k_{n,m}\|_2, \qquad \ell^{\text{smpl}}_{k,n,m} = \|\mathcal{J}^{\text{2D,head}}_{n,m} - \mathcal{J}^{\text{2D},k}_{n,m}\|_2 \tag{2}$$

$$r = \sum_{(k,n,m)\in\mathcal{S}} w^k_{n,m} \frac{\ell^{\text{smpl}}_{k,n,m}}{\ell^{\text{img}}_{k,n,m}} \tag{3}$$

where $\mathcal{J}^{\text{2D}}$ denotes the projected SMPL joint and $\mathcal{S}$ the set of joint-frame-person triplets used for calibration. The weight $w^k_{n,m}$ is the product of the predicted visibility and confidence, normalized to sum to one over $\mathcal{S}$, which makes unreliable joints contribute less to the ratio. Finally, we obtain the calibrated scale as $s^* = r \cdot s$, enabling consistency between the scene and the SMPL meshes.

3.3 Multi-View Fusion Module

For multi-view human-scene reconstruction, existing approaches (Müller et al., 2025; Rojas et al., 2025; Liu et al., 2026a) often rely on test-time optimization or precomputed cross-view correspondences, which are required for human and scene alignment. In contrast, our bidirectional anchoring already aligns humans and scene within each view, so per-view human estimates can be merged directly in a shared world coordinate system. Building on this, we introduce a *multi-view fusion module* that fuses per-view human estimates into a coherent global representation without test-time optimization or precomputed cross-view correspondences.

Component-wise Aggregation. The predicted human representation contains both view-invariant and view-dependent components. For each human, rather than uniformly averaging all per-view estimates, we aggregate each component according to its geometric properties. The canonical pose $\theta^v_{t,m}$ and body shape $\beta^v_{t,m}$ are shared across views, so we compute their weighted averages $(\theta^{\text{avg}}_{t,m}, \beta^{\text{avg}}_{t,m})$ using reliability weights derived from the joint visibility and confidence estimates in Sec. 3.2, which reduces sensitivity to occlusions. In contrast, the root rotation $R^v_{t,m}$ and 3D head translation $\tau^v_{t,m}$ are defined in each camera coordinate system, so we transform them into a shared world coordinate system. For the global root rotation $\bar{R}_{t,m}$, we compute a weighted average of the 6D representations and map the result back to $SO(3)$. For the global head translation $\bar{\tau}_{t,m}$, we observe that using multi-view ray triangulation instead of averaging translation enforces geometric consistency across views. Together, our component-wise aggregation strategies improve multi-view pose accuracy.

Refine Transformer. Component-wise aggregation provides robust initial estimates, but it does not fully exploit complementary information across views. We therefore introduce a *refine transformer* that refines the averaged pose and shape estimates (Fig. 4). For each person m at time t, we form a token sequence consisting of one learnable multi-view query Q^{MV} and the corresponding refined per-view queries $\hat{Q}^v_{t,m}$ from Sec. 3.1. We add the embedding of the averaged and per-view parameters to the sequence and obtain the input sequence $[X^{\text{MV}}_{t,m}, X^1_{t,m}, \ldots, X^V_{t,m}]$ as

$$X^{\text{MV}}_{t,m} = Q^{\text{MV}} + \text{Enc}([\theta^{\text{avg}}_{t,m} | \beta^{\text{avg}}_{t,m}]), \tag{4}$$

$$X^v_{t,m} = \text{Proj}(\hat{Q}^v_{t,m}) + \text{Enc}([\theta^v_{t,m} | \beta^v_{t,m}]). \tag{5}$$

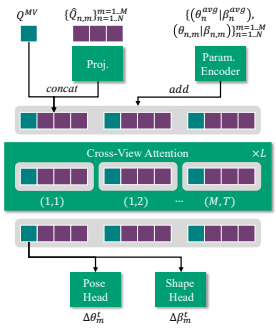

Figure 4: Refine TF.

The shared parameter encoder $\text{Enc}(\cdot)$ maps concatenated pose and shape parameters to the token space, and $\text{Proj}(\cdot)$ linearly projects each refined per-view query to the same dimension. The resulting sequence is processed by the refine transformer, where self-attention operates across the views of the same person at the same timestep. The multi-view output token of each sequence is decoded into pose and shape residuals, and the final pose and shape $(\bar{\theta}_{t,m}, \bar{\beta}_m)$ are obtained as

$$\bar{\theta}_{t,m} = \theta^{\text{avg}}_{t,m} + \Delta\theta_{t,m}, \qquad \bar{\beta}_m = \frac{1}{T}\sum_{t=1}^{T} \left(\beta^{\text{avg}}_{t,m} + \Delta\beta_{t,m}\right). \tag{6}$$

Geometric Multi-Person Association. Previous approaches rely on manually assigned identities (Müller et al., 2025), which increase system complexity, or appearance-based ReID (Li et al., 2024), which is unreliable for visually similar individuals. We instead associate people using pose and global 3D position. Within each view, we form tracklets by matching refined query tokens $\hat{Q}^v_{t,m}$ across consecutive timesteps based on ℓ_2 distance, and keep unmatched tokens in memory for re-tracking. We construct a 2-nearest-neighbor view graph using mean camera centers. For each connected view pair, we compute matching costs between all tracklet pairs as

$$\mathcal{C}(a,b) = \frac{1}{|\mathcal{T}_{a,b}|} \sum_{t \in \mathcal{T}_{a,b}} \frac{1}{J} \sum_{j=1}^{J} \left(\lambda_p \left\| \mathcal{J}^{\text{3D},a}_{t,j} - \mathcal{J}^{\text{3D},b}_{t,j} \right\|_2 + \lambda_c \left\| \mathcal{J}^{\text{3D},a,\text{canon}}_{t,j} - \mathcal{J}^{\text{3D},b,\text{canon}}_{t,j} \right\|_2 \right). \tag{7}$$

Table 1: Quantitative comparison on RICH and EMDB-2 datasets. Ours-multi uses all available views and is reported for reference. red : the best, yellow : the second best.

	Settings			RICH			EMDB-2		
	Multi-View	Multi-Person	Scene	WA-MPJPE↓	W-MPJPE↓	RTE(%)↓	WA-MPJPE↓	W-MPJPE↓	RTE(%)↓
JOSH3R	✗	✓	✓	-	-	-	220.0	661.7	13.1
UniSH	✗	✗	✓	118.1	183.2	4.8	118.5	270.1	5.8
Human3R	✗	✓	✓	110.0	184.9	3.3	112.2	267.9	2.2
Ours-mono	✗	✓	✓	84.9	132.4	2.6	83.6	214.4	1.4
Ours-multi	✓	✓	✓	**45.4**	**61.2**	**0.9**	-	-	-

Table 2: Quantitative comparison on EgoExo4D and EgoHumans datasets. For fair comparison, all methods are evaluated with ground-truth identities. All metrics are in meters.

	Settings			EgoExo4D		EgoHumans			
	Video	No ReID	No Optim.	W-MPJPE† ↓	PA-MPJPE↓	W-MPJPE† ↓	GA-MPJPE↓	PA-MPJPE↓	Runtime↓
HSfM (init.)	✗	✗	✓	5.29	0.07	4.28	0.51	0.06	∼21s
HSfM	✗	✗	✗	0.56	0.06	1.04	0.21	0.05	∼118s
HAMSt3R	✗	✓	△	0.51	0.09	3.80	0.42	0.14	∼32s
Ours	✓	✓	✓	0.23	0.06	0.37	0.20	0.05	∼1.1s

where $\mathcal{T}_{a,b}$ denotes the set of overlapping timesteps, and $\mathcal{J}^{\text{3D,canon}}$ denotes joints without root rotation or translation. We set $\lambda_c = 0.8$ and $\lambda_p = 0.2$. For each view-graph edge, we use Hungarian matching to associate tracklets. We merge accepted matches across pairs using Union-Find, obtaining globally consistent identities. Unlike other components, our association is not learnable and runs only at inference, without updating any camera or human parameters. Further details are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

We train HARMONI using BEDLAM (Black et al., 2023), 3DPW (Von Marcard et al., 2018), MPII (Andriluka et al., 2014), MSCOCO (Lin et al., 2014), and MammaSyn (Cuevas Velasquez et al., 2026) datasets. We evaluate our framework on two conventional tasks: (i) global human motion estimation and (ii) multi-view human pose estimation. We compare against methods with publicly available implementations or directly comparable results reported under the same established evaluation protocols, including Human3R (Chen et al., 2026) and HSfM (Müller et al., 2025). For global human motion estimation, we evaluate on EMDB-2 (Kaufmann et al., 2023), which contains monocular videos captured by dynamic cameras, and RICH (Huang et al., 2022), which provides fixed multi-view recordings. Following the evaluation protocol of Human3R (Chen et al., 2026), we report WA-MPJPE and W-MPJPE, which measure mean joint errors after aligning each 100-frame segment using the entire segment and only its first two frames, respectively, and RTE, which measures the normalized root translation error after SE(3) alignment. For multi-view human pose estimation, we evaluate on EgoHumans (Khirodkar et al., 2023) and EgoExo4D (Grauman et al., 2024) under the single-timestep evaluation protocol of HSfM (Müller et al., 2025). We report W-MPJPE† using camera-only SE(3) alignment, GA-MPJPE after jointly aligning all people, and PA-MPJPE after person-wise Procrustes alignment. Implementation details are provided in Appendix B.

4.2 Human and Scene Reconstruction

Global Motion Estimation. We compare our framework with feed-forward monocular human-scene reconstruction methods on EMDB-2 and RICH. Among the baselines, JOSH3R (Liu et al., 2026b) requires 2D human masks, UniSH (Li et al., 2026) relies on bounding-box detection and supports only a single person, and Human3R (Chen et al., 2026) supports multiple people but only monocular inputs. As shown in Table 1, our framework achieves lower global motion errors than these baselines in the monocular setting, while using multiple views further improves accuracy.

Table 3: Ablation on multi-view fusion strategies.

RICH-MV	WA-M.↓	W-M.↓	RTE(%)↓	PA-M.↓
Ours-mono	84.9	132.4	2.6	51.0
(1) Avg.	65.0	93.3	1.2	38.2
(2) Tri.+Avg.	48.2	67.0	1.0	38.2
(3) Tri.+Conf.	47.8	66.2	1.0	37.4
(4) Tri.+Conf.+Refine	**45.4**	**61.2**	**0.9**	**35.6**

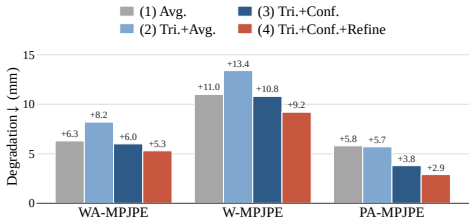

Figure 5: Sensitivity analysis.

Multi-View Pose Estimation. We compare our framework with existing multi-view human-scene reconstruction methods under the single-timestep protocol used by HSfM (Müller et al., 2025) and HAMSt3R (Rojas et al., 2025). We do not include Trophies (Liu et al., 2026a), as its multi-view evaluation follows a different protocol and its implementation is not publicly available. Unlike ours, Trophies relies on test-time optimization with external metric depth, 2D keypoints, and human masks, and does not address cross-view identity association. As shown in Table 2, our method reduces global pose error while achieving competitive local pose accuracy without test-time optimization. On a single EgoHumans scene (three people and four views) using one NVIDIA V100 GPU, it runs over $100\times$ faster than HSfM and $28\times$ faster than HAMSt3R. Finally, we compare applying the calibration ratio r to the scene scale (Ours-s^*) against keeping the scene scale fixed and instead rescaling the SMPL meshes (Ours-s). We report W-MPJPE† and mean camera translation error (TE), both computed after only a rigid SE(3) alignment. As both metrics are sensitive to metric scale, Table 4 shows that calibrating the scene using the human prior improves the estimated scene scale.

Table 4: Ablation on scale calibration. Both metrics are computed in meters.

	EgoExo4D		EgoHumans	
	W-M.† ↓	TE.↓	W-M.† ↓	TE.↓
Ours-s	0.31	0.54	1.56	3.01
Ours-s^*	**0.23**	**0.19**	**0.37**	**0.31**

Scene Reconstruction. We measure camera accuracy on EMDB-2 with the absolute trajectory error (ATE) and the relative pose error (RPE$_t$), and depth accuracy on Bonn (Palazzolo et al., 2019) with the absolute relative error (AbsRel) and the thresholded accuracy ($\delta < 1.25$). As shown in Table 5, our framework preserves the accuracy of the scene prior, whereas UniSH (Li et al., 2026), which fine-tunes its backbone, degrades camera accuracy. This supports our frozen scene design choice in Sec. 3.1.

Table 5: Camera pose and depth accuracy.

	Camera (EMDB-2)		Depth (Bonn)	
	ATE↓	RPE$_t$ ↓	AbsRel↓	δ<1.25 ↑
CUT3R	.059	.013	.048	.970
Human3R	.037	.012	.050	.961
Pi3	**.025**	.010	**.026**	.985
UniSH	.121	.021	.034	.985
Pi3X	.026	**.008**	**.026**	**.986**
Ours	.026	**.008**	**.026**	**.986**

4.3 Analysis

Fusion Components. We ablate multi-view fusion components on RICH using PA-MPJPE to evaluate local human quality. As shown in Table 3, uniform averaging (Avg.) performs worst, whereas replacing translation averaging with ray triangulation (Tri.) yields the largest gain. Confidence weighting (Conf.) further improves accuracy, and the refine transformer (Refine) enhances local pose. Fig. 5 evaluates robustness to unreliable views by occluding a person of a randomly selected view by 60% with a gray box, as in Kocabas et al. (2021). Triangulation alone is most sensitive to such views, while confidence weighting suppresses unreliable views and reduces the degradation. The refine transformer further lowers it, validating our fusion design choice.

Bidirectional Anchoring. We evaluate the effect of scale and depth alignment in bidirectional anchoring on EMDB-2 (Table 6). Without scale calibration, errors increase substantially due to the mismatch between two priors. For depth alignment, direct depth performs worst, while direct translation shows better accuracy. Our depth residual prediction performs best.

Table 6: Ablation on alignment strategies. All variants are trained up to Stage 1.

EMDB-2	WA-M.↓	W-M.↓	RTE(%)↓
w/o scale calib.	157.8	422.3	4.2
Direct transl.	100.9	251.4	1.9
Direct depth	111.8	276.4	2.2
Ours-Stage 1	**84.7**	**216.4**	**1.4**

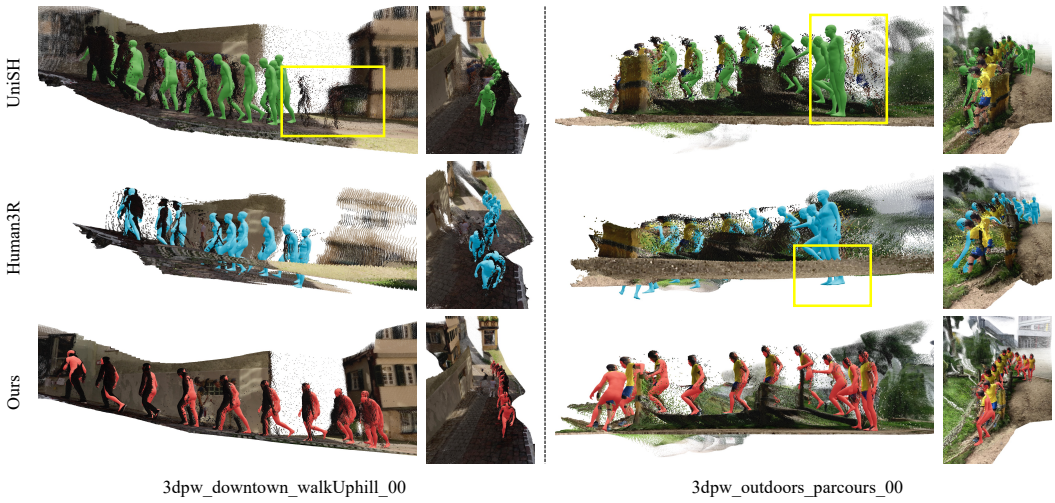

Figure 6: **Qualitative results on the 3DPW dataset.** We compare our method with monocular human-scene reconstruction baselines, including UniSH and Human3R.

Association Ablation. We evaluate multi-person association on EgoHumans using ReID accuracy, precision, and recall in Table 7. As reported in Müller et al. (2025), appearance-based ReID achieves 51.2% accuracy, showing the difficulty of distinguishing similar individuals. Our geometric association substantially improves accuracy to 90.9%. Pose alone yields the highest recall but the lowest precision by over-matching similar poses. Combining pose and position performs best, with $\lambda_c = 0.8$ and $\lambda_p = 0.2$, while accuracy remains stable for $\lambda_c \in [0.2, 0.8]$.

Table 7: Ablation on association weights.

λ_c	λ_p	Acc.(%)	F1(%)	Prec.(%)	Recall(%)
0.0	1.0	89.5	78.9	**94.1**	67.9
0.2	0.8	89.9	79.9	94.0	69.6
0.5	0.5	90.7	82.1	92.8	73.6
0.8	0.2	**90.9**	**83.3**	88.5	78.7
1.0	0.0	68.3	59.6	47.2	**80.8**

Qualitative Analysis. We compare qualitative results with other monocular human–scene reconstruction methods on the 3DPW dataset. As shown in Fig. 6, UniSH shows depth misalignment between the scene point cloud and human mesh. In addition, Human3R produces penetrations when the estimated scene and human scales are inconsistent. In contrast to these baselines, our method achieves coherent human–scene alignment. Additional results are provided in Appendices A and C.

5 Conclusion

In this paper, we present HARMONI, a feed-forward framework that reconstructs cameras, scene, and humans with identities from monocular or multi-view video without test-time optimization. We introduce *bidirectional anchoring* to bridge complementary human and scene priors: human keypoints calibrate scene scale to match the human body, while scene depth guides global human placement. Building on the per-view alignment, our *multi-view fusion module* combines confidence-weighted averaging, triangulation, a refine transformer, and geometry-based association to handle unreliable views and visually similar individuals. Experiments demonstrate improved global motion and multi-view pose accuracy by up to 28% and 64% over previous human-scene reconstruction methods, while running over $28\times$ faster than optimization-based multi-view approaches.

Limitations and Future Work. HARMONI shares a common limitation with the CenterNet-based paradigm in that it cannot recover when the head is fully invisible in the monocular setting. In addition, since scale calibration uses the human body, it may be biased for people whose body sizes deviate from the training distribution, such as children. We plan to extend the work by distilling heterogeneous priors into a unified encoder.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the method (writing and debugging parts of the model, training, and evaluation code), to provide feedback on the design of experiments, and to help interpret experimental results. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or assist in writing proofs, propose or refine hypotheses, or clean and reformat datasets; translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper, to edit the manuscript for grammar and readability, to suggest the paper's structure, title, and keywords, to search for and summarize related literature and to verify and format references, to suggest experimental parameters, and to write auxiliary code such as visualization and data-processing scripts. We have reviewed all AI-assisted work. All AI-assisted code was reviewed and tested for correctness by the authors, and all reported results were produced by running the authors' verified code. AI feedback on experimental design and AI-suggested interpretations of results were critically assessed by the authors, who made all final decisions on the experiments and the conclusions drawn from them. AI-drafted text was rewritten and checked by the authors, and all AI-identified literature was verified against the original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve the collection of new human-subject data; all experiments use publicly available datasets under their respective licenses. Since our method reconstructs and tracks multiple people across cameras, it could potentially be misused for surveillance without consent. We encourage its use only with the informed consent of captured individuals and in compliance with applicable privacy regulations.

REPRODUCIBILITY STATEMENT

The method is described in Sec. 3, and full details on architecture, training schedule, losses, hyperparameters, and evaluation protocols are provided in Appendix B. All datasets used are publicly available. We will release our code and pretrained models upon publication.

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.

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*, pp. 3686–3693, 2014.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024.

Paul Beardsley, Phil Torr, and Andrew Zisserman. 3d model acquisition from extended image sequences. In *European conference on computer vision*, pp. 683–695. Springer, 1996.

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*, pp. 8726–8737, 2023.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *International Conference on Learning Representations*, 2026.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025.

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)*, 2026.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

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*, pp. 281–299. Springer, 2026.

Andrew W Fitzgibbon and Andrew Zisserman. Automatic camera recovery for closed or open image sequences. In *European conference on computer vision*, pp. 311–326. Springer, 1998.

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

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*, pp. 14783–14794, 2023.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

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*, pp. 13274–13285, 2022.

Minhyuk Hwang, Sangmin Kim, Seunguk Do, Daneul Kim, and Jaesik Park. 4d human-scene reconstruction from low-overlap captures. In *ACM SIGGRAPH 2026 Conference Proceedings*, 2026.

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*, pp. 14632–14643, 2023.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19807–19819, 2023.

Sangmin Kim, Seunguk Do, and Jaesik Park. Showmak3r: Compositional tv show reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 864–874, 2025.

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

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, et al. Unish: Unifying scene and human reconstruction in a feed-forward pass. *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yu-Jhe Li, Yan Xu, Rawal Khirodkar, Jinhyung Park, and Kris Kitani. Multi-person 3d pose estimation from multi-view uncalibrated depth cameras. *arXiv preprint arXiv:2401.15616*, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

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

Jinpeng Liu, Yukang Xu, Yutong Li, and Xingyu Liu. Trophies: Temporal reconstruction of places, humans, and cameras from multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

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

H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2299–2307, 2022.

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21948–21958, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. 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`.

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

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

YM Pradyumna, Yuxuan Xue, Yue Chen, Nikita Kister, István Sárándi, and Gerard Pons-Moll. Graft: Geometric refinement and fitting transformer for human scene reconstruction. In *European Conference on Computer Vision*, pp. 250–267. Springer, 2026.

Sara Rojas, Matthieu Armando, Bernard Ghanem, Philippe Weinzaepfel, Vincent Leroy, and Gregory Rogez. Hamst3r: Human-aware multi-view stereo 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5027–5037, 2025.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pp. 501–518. Springer, 2016.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pp. 519–528. IEEE, 2006.

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.

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*, pp. 2070–2080, 2024.

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*, pp. 21132–21142, 2026.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in neural information processing systems*, 36:39033–39051, 2023.

Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992.

Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael J. Black, Daniel Holden, and Carsten Stoll. HUMOS: Human motion model conditioned on body shape. In *European Conference on Computer Vision*, pp. 133–152. Springer, 2024.

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)*, pp. 601–617, 2018.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 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*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5261–5271, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024a.

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. In *International Conference on Learning Representations*, 2026b.

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*, pp. 467–487. Springer, 2024b.

Zihan Wang, Jiashun Wang, Jeff Tan, Yiwen Zhao, Jessica Hodgins, Shubham Tulsiani, and Deva Ramanan. Contact-guided real2sim from monocular video with planar scene primitives. In *International Conference on Learning Representations*, volume 2026, pp. 115452–115468, 2026c.

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*, pp. 21222–21232, 2023.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pp. 180–200. Springer, 2022.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Appendix

In this section, we provide the implementation details of our framework, HARMONI, along with the additional analyses and results to support our claims. In Sec. A, we present additional analyses of our framework. In Sec. B, we describe the implementation details. In Sec. C, we present additional results across diverse scenarios, as well as failure cases.

A Additional Analysis

A.1 Bidirectional Anchoring

To further analyze the effect of our bidirectional anchoring, we perform physical plausibility experiment, trajectory comparison and visualize heatmaps on truncated and occluded scenarios.

Physical Plausibility. We measure physical plausibility on RICH (Huang et al., 2022), where calibrated cameras and scene scans are provided. We define contact vertices as ground-truth body vertices within 3 cm of the scan. Each predicted body is transformed into the scan frame using the ground-truth camera pose, and the corresponding vertices are compared with the local contact surface. Following HUMOS (Tripathi et al., 2024), we report penetration and float as their mean distances (cm) below and above the surface. As shown in Table 8, our bidirectional anchoring achieves lower penetration and float than UniSH (Li et al., 2026) and Human3R (Chen et al., 2026), which demonstrates its effectiveness in placing humans coherently within the scene.

Table 8: Physical plausibility evaluation.

RICH	Penet.↓	Float↓
UniSH	13.7	11.5
Human3R	5.3	18.1
Ours	**5.0**	**7.6**

Trajectory Comparison. In Fig. 7, we compare our global root trajectories with Human3R, the previous state-of-the-art method. We select four EMDB-2 sequences with long walking paths, and visualize from a top-down view with the ground truth. Each trajectory is aligned to the ground truth once, using its first 100 frames. Human3R gradually drifts away from the ground truth as the sequence progresses, whereas our trajectories closely follow it in both shape and scale over the entire sequence. This shows that bidirectional anchoring yields globally consistent human motion.

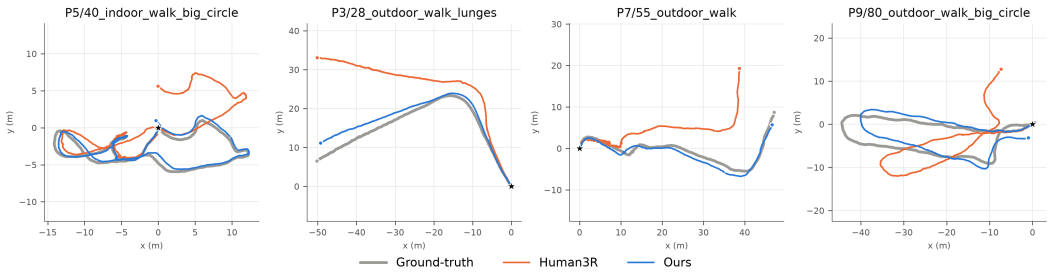

Figure 7: Trajectory comparison.

Heatmap Visualization. We analyze how our bidirectional anchoring behaves under truncation and occlusion by visualizing the 2D keypoints and the heatmaps of the spine and both feet. When the lower body is truncated by the image boundary (Fig. 8), the predicted visibility of both feet converges to zero. When the feet are inside the image but occluded (Fig. 9), the visibility remains close to 1.0 while the confidence decreases. Since the calibration weight $w_{n,m}^{k}$ in Eq. 3 is defined as the product of visibility and confidence, truncated and occluded joints contribute less to the calibration ratio, which allows the scene scale to be calibrated robustly in both cases. As shown in both Fig. 8 and Fig. 9, both humans are placed coherently in the scene.



Figure 8: Heatmap visualization of truncated scene and the reconstructed output.



Figure 9: Heatmap visualization of occluded scene and the reconstructed output.

A.2 Robustness Analysis

We present qualitative results of our framework across challenging scenarios, including in-the-wild scenes with truncated, occluded, or crowded subjects, multi-view settings, and monocular sequences with temporal occlusions.

In-the-wild Scenarios. We show qualitative results in challenging in-the-wild scenarios in Fig. 10 and Fig. 11. Fig. 10 includes individuals with truncation (yellow box), occlusion (purple box), and objects placed in front of the head (cyan box). Our framework is able to recover the truncated and occluded body parts, while robustly aligning the humans with the scene. Also, our framework is able to infer correct depth even when the head is partially covered by an object. In Fig. 11, we increase the maximum number of people M from 10 to 30 at inference, and reconstruct the crowded scene. Although trained with at most 10 people per frame, our framework successfully recovers more than 20 people. Moreover, our depth residual prediction preserves the relative depth ordering among people.

Multi-View Scenario. We illustrate how per-view estimates are merged into the global representation in multi-view scenario of EgoExo4D (Grauman et al., 2024). In Fig. 12, we show the multi-view reconstruction process of a scene with 3 people and 4 views. Our framework is able to reconstruct global humans even when detection fails in View 4 (yellow box). Furthermore, two people are observed only in View 3 (pink box) and are still successfully placed in the world coordinate system even though they are not visible in any other view. This demonstrates the practical applicability of our framework, which requires only a single visible view.

Temporal Occlusion in Monocular Setting. Since our detection follows the CenterNet-based paradigm (Zhou et al., 2019), a person whose head is fully occluded cannot be detected. In the multi-view setting, people can be recovered with other visible views. However in a monocular set-

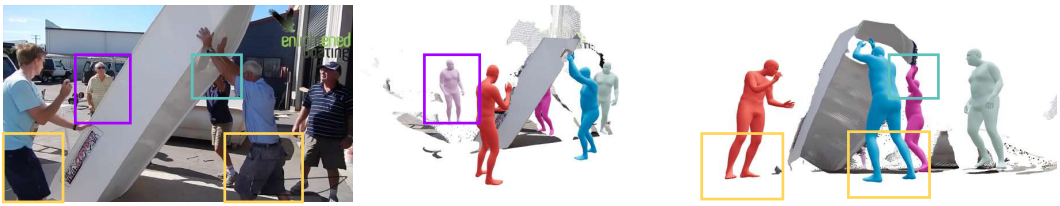

Figure 10: Challenging scenario with truncation, occlusion, and object in front of the head.



Figure 11: Challenging scenario with crowded scene.

ting, when the head is fully invisible, it cannot be recovered. For practicality, we track people using their refined query features and keep unmatched tracks in memory so that they can be re-matched after temporary missed detections. As shown in Fig. 13, our framework consistently tracks people through occlusion without identity switches.

A.3 Additional Experiments

Finally, we compute runtime per component, compare with other human motion recovery methods, and analyze the effect of our scale calibration on another baseline.

Runtime Per Component. To further analyze the runtime cost in Table 2, we measure the computation time of each module of our framework in Table 9. We select the same scene with three people and four views from EgoHumans (Khirodkar et al., 2023), repeat the inference five times, and report the average runtime on a single NVIDIA V100 GPU.

Table 9: Runtime analysis on multi-view scene.

Module	Runtime ↓
Scene encoder	0.09s
Scene decoder	0.25s
Scene heads	0.23s
Human encoder	0.40s
Human decoder	0.01s
SMPL head	0.02s
Association	0.05s
Multi-View fusion	0.08s
Total	**1.13s**

Comparison with Human Motion Recovery Methods. We report global motion quality with world-grounded human motion recovery methods (Ye et al., 2023; Shin et al., 2024; Shen et al., 2024; Wang et al., 2024b) on EMDB-2 (Kaufmann et al., 2023) in Table 10. These methods estimate only human motion and rely on external SLAM systems such as DROID-SLAM (Teed & Deng, 2021) or DPVO (Teed et al., 2023), whereas our framework jointly reconstructs cameras, scene, and humans in a feed-forward pass. Our method achieves competitive global motion accuracy.

Table 10: Comparison with world-grounded human motion recovery methods on EMDB-2.

	Settings		EMDB-2		
	Scene	Cam	WA-MPJPE↓	W-MPJPE↓	RTE(%)↓
SLAHMR	✗	DROID-SLAM	326.9	776.1	10.2
WHAM	✗	DPVO	135.6	354.8	6.0
GVHMR	✗	DPVO	111.0	276.5	2.0
TRAM	✗	DROID-SLAM	**76.4**	222.4	**1.4**
Ours	✓	Feed-Forward	83.6	**214.4**	**1.4**



Figure 12: Multi-view reconstruction process.

Scale Calibration on Human3R. To examine whether scale calibration transfers to other frameworks, we apply it to Human3R by computing the ratio r in Eq. 3 from its SMPL meshes and our 2D keypoints. We do not apply depth residual prediction, as it requires re-training. As shown in Table 11, calibration improves global motion accuracy on EMDB-2 but slightly degrades it on RICH, as Human3R's translation is not anchored on the scene depth. This suggests that scale calibration is most effective when combined with depth anchoring.

Table 11: Scale calibration applied to Human3R without depth residual prediction.

	RICH			EMDB-2		
	WA-MPJPE↓	W-MPJPE↓	RTE(%)↓	WA-MPJPE↓	W-MPJPE↓	RTE(%)↓
Human3R	**110.0**	**184.9**	**3.3**	112.2	267.9	2.2
Human3R + Calib.	111.2	190.1	3.5	**106.4**	**250.4**	**2.1**

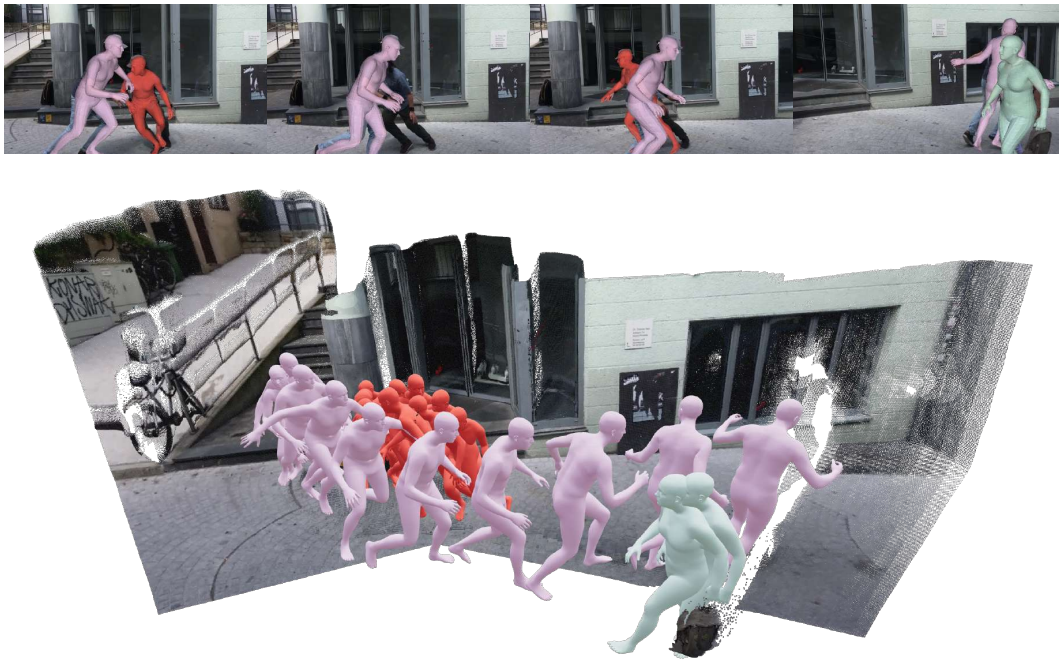

Figure 13: Temporal occlusion example.

B Implementation Details

B.1 Model Architecture

Network Design. We freeze all Pi3X (Wang et al., 2026b) modules and the DINOv2 (Oquab et al., 2024) ViT backbone from Multi-HMR (Baradel et al., 2024), including the detection MLP that predicts a person-center score map and the offset MLP that localizes each human head. For human reconstruction, we train a mask MLP to predict full-resolution person masks and a human decoder to refine the detected person tokens. The decoder consists of transformer blocks with self-attention among the person tokens, cross-attention to the image features, and a feed-forward network. Three SMPL heads decode each refined token into full body pose, shape coefficients, and a depth residual. We also train heatmap heads implemented as two linear projections, one applied to the refined person query and the other to the human feature map, whose dot products produce per-joint heatmaps decoded using a differentiable soft-argmax. Two additional linear heads predict the per-joint visibility and confidence logits used in Sec. 3.2. For multi-view fusion, we train a shared parameter encoder, implemented as a two-layer MLP with GELU activation, to map the concatenated 6D pose and shape parameters to the transformer width. The refine transformer consists of transformer layers with multi-head self-attention and GELU feed-forward networks. Two MLP heads decode the fused token into pose and shape residuals. Their final layers are zero-initialized so that the module initially preserves the averaged parameters.

Association Algorithm. Next, we describe the multi-person association algorithm in Alg. 1. Given the refined human queries $\hat{Q}$ and SMPL 3D joints $\mathcal{J}^{3D}$ from Sec. 3.1, we first track humans within each view by matching the queries of the current detections to a track memory containing the stored query of each observed track. Matching is performed using pairwise L2 distances and Sinkhorn optimal transport. Matched tracks update their stored queries, while unmatched tracks remain in memory so that they can be re-matched after temporary misses. We then filter outliers based on the average displacement of the world-space 3D joints between consecutive observations of each tracklet. After per-view tracking, we associate human tracklets across views by applying the Hungarian algorithm to the matching cost defined in Eq. 7. Finally, the matched tracklets are merged using union-find, and global human identities are assigned.

Algorithm 1 Multi-Person Association

1. Per-View Tracking
for each view $v = 1, \ldots, V$ **do**
  Initialize human track IDs and the track memory with all detections at $t = 1$
  for frame $t = 2, \ldots, T$ **do**
    Compute pairwise L2 distance matrix $\mathcal{D}$ between tokens in the track memory and detections at t
    Perform one-to-one matching via Sinkhorn optimal transport
    Assign new human track IDs to unmatched detections
    Update the track memory with matched tokens; keep unmatched tracks for re-matching
  end for
  for each human tracklet m **do**
    Transform 3D joints of tracklet m into world coordinates using camera extrinsics
    for each observed frame t of tracklet m with previous observed frame t' **do**
      Compute mean joint displacement $\Delta = \frac{1}{J} \sum_j \|\mathcal{J}^m_{t,j} - \mathcal{J}^m_{t',j}\|_2$
      if $\Delta > \gamma_{\text{outlier}}$ **then**
        Remove frame t from human tracklet m
      end if
    end for
  end for
end for

2. Cross-View Association
Compute mean camera center for each view
Construct view connectivity graph using k-NN
for each connected view pair (v_i, v_j) **do**
  for each human tracklet pair (a, b) **do**
    Compute matching cost $\mathcal{C}(a, b)$ using Eq. 7
  end for
  Solve one-to-one assignment via Hungarian algorithm
  Retain matches with $\mathcal{C} < \gamma_{\text{reid}}$
end for
Merge retained matches across all view pairs via union-find
Assign one global human ID per connected component

B.2 Training Setup

We train our framework in three stages on 4 NVIDIA A6000 GPUs. Each stage takes approximately one day, resulting in a total training time of about three days. Throughout training, we set the maximum number of people per view to 10.

Training Schedule. In Stage 1, we freeze the scene encoder, scene decoder, and scene output heads, as well as the human encoder and the detection and offset MLPs. We train the human decoder, SMPL head, mask MLP, heatmap heads, and visibility and confidence heads for 20 epochs. Scale calibration is disabled for the first 10 epochs to focus on learning human shape, pose, and translation, and is enabled for the remaining 10 epochs. For each training sample, the input sequence length is randomly sampled between 2 and 24 frames. We use the AdamW optimizer with a OneCycleLR scheduler and an initial learning rate of 5×10^{-5}. Ground-truth depth maps, camera parameters, and head keypoint locations are provided as input condition during this stage. The multi-view fusion module is disabled.

In Stage 2, we additionally freeze the human decoder, SMPL head, and mask MLP, and train only the heatmap, visibility, and confidence heads for 20 epochs. The model takes a single image as input, with only the head keypoint locations provided, while the camera parameters are predicted by the scene branch. We use the AdamW optimizer with a OneCycleLR scheduler and an initial learning rate of 1×10^{-4}. The multi-view fusion module remains disabled.

In Stage 3, we freeze all modules except the multi-view fusion module and train it for 20 epochs. Specifically, we train the shared parameter encoder, refine transformer, and pose and shape residual heads. Each training sample contains a single time step with 2 to 8 views, and ground-truth person identities across views are provided. Ground-truth head keypoint locations and camera intrinsics are provided in every batch, whereas ground-truth camera extrinsics are provided in a randomly selected

50% of the batches. We use the AdamW optimizer with a OneCycleLR scheduler, starting with a learning rate of 1×10^{-4}. To improve robustness to unreliable views, we augment the lower body of a randomly selected person in one randomly selected view by pasting Pascal VOC (Everingham et al., 2010) objects to simulate occlusion and by applying a random crop to simulate truncation. The two augmentations are applied independently, and the probability of each increases linearly from 0 to 0.5 over the first 5 epochs. At inference, all camera parameters are predicted by the scene branch without ground-truth extrinsics.

Dataset Configuration. Across all stages, we use dynamic-resolution training with aspect ratios ranging from 0.5 to 2.0. We augment the images using random cropping, focal-length jitter, color jitter, and blurring. Stage 1 is trained only on BEDLAM (Black et al., 2023), a synthetic dataset that provides SMPL-X parameters, depth maps, and camera parameters for every frame. We compute per-joint occlusion labels for the seven detector joints by depth-testing the projected SMPL-X joints against the rendered depth maps. To improve performance on in-the-wild scenarios, Stage 2 is trained on single images from BEDLAM, 3DPW (Von Marcard et al., 2018), MPII (Andriluka et al., 2014), and MSCOCO (Lin et al., 2014), sampled at a ratio of 1:1:1:3. For the three in-the-wild datasets, we use SMPL-X pseudo-ground truth from NeuralAnnot (Moon et al., 2022) and precompute the 2D projections of the joints and vertices to supervise the 2D joint loss. Occlusion labels are obtained from OpenPose (Cao et al., 2019) confidence scores for 3DPW, keypoint visibility flags for MSCOCO, and visibility annotations for MPII. Joints without an available label are excluded from the confidence loss. For Stage 3, we train the multi-view fusion module on MammaSyn (Cuevas Velasquez et al., 2026), a synthetic dataset containing approximately 895 two-person interaction sequences rendered from eight synchronized cameras, with world-space SMPL-X parameters and cross-view-consistent person identities. Per-joint occlusion labels are derived from the per-view vertex visibility provided by the dataset.

Loss Design. In Stage 1, we use geometric, parametric, and detection losses. The geometric losses $\mathcal{L}_{\text{v3d}}$ and $\mathcal{L}_{\text{j3d}}$ supervise the predicted SMPL vertices and 3D joints in both camera and pelvis-relative coordinates, after aligning the predicted translation and scale with the ground truth using the scale factor s_{opt} from the ROE solver (Wang et al., 2025c). Their 2D projections are supervised by $\mathcal{L}_{\text{v2d}}$ and $\mathcal{L}_{\text{j2d}}$. The parametric losses supervise the SMPL pose θ^{full}, shape β, and translation τ, and since scale calibration depends on body size, an L1 loss $\mathcal{L}_{\text{height}}$ additionally supervises the rest-pose body height derived from β. The detection losses consist of a BCE mask loss $\mathcal{L}_{\text{mask}}$, an image-size-normalized L1 loss $\mathcal{L}_{\text{loc}}$ on in-frame detector joints, and BCE losses $\mathcal{L}_{\text{vis}}$ and $\mathcal{L}_{\text{conf}}$ on the visibility and confidence logits, where $\mathcal{L}_{\text{conf}}$ is applied only to joints with occlusion labels:

$$\begin{aligned}\mathcal{L}_{\text{stage1}} = \lambda_{\text{v3d}}\mathcal{L}_{\text{v3d}} + \lambda_{\text{j3d}}\mathcal{L}_{\text{j3d}} + \lambda_{\text{v2d}}\mathcal{L}_{\text{v2d}} + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}} + \lambda_{\theta}\mathcal{L}_{\theta^{\text{full}}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\tau}\mathcal{L}_{\tau} \\ + \lambda_{\text{height}}\mathcal{L}_{\text{height}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}},\end{aligned} \tag{8}$$

where $\lambda_{\text{v3d}} = \lambda_{\text{j3d}} = \lambda_{\tau} = 5.0$ and all other weights are 1.0.

In Stage 2, since the in-the-wild datasets lack ground-truth cameras, the 2D reprojection losses are computed with the predicted intrinsics and without s_{opt}, and are combined with the joint detection losses:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{v2d}}\mathcal{L}_{\text{v2d}} + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}, \tag{9}$$

with all weights set to 1.0.

In Stage 3, we supervise the fused SMPL parameters from the multi-view fusion module. The fused root rotation and translation are transformed into each input camera frame and passed through the SMPL-X layer. We apply $\mathcal{L}_{\theta^{\text{full}}}$, $\mathcal{L}_{\beta}$, $\mathcal{L}_{\text{j2d}}$, and $\mathcal{L}_{\text{v2d}}$, and since the fused body is anchored at the triangulated head, $\mathcal{L}_{\text{j3d}}$ and $\mathcal{L}_{\text{v3d}}$ are computed in head-relative coordinates. An L2 penalty $\mathcal{L}_{\Delta}$ regularizes the pose and shape residuals:

$$\begin{aligned}\mathcal{L}_{\text{stage3}} = \lambda_{\text{v3d}}\mathcal{L}_{\text{v3d}} + \lambda_{\text{j3d}}\mathcal{L}_{\text{j3d}} + \lambda_{\text{v2d}}\mathcal{L}_{\text{v2d}} + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}} \\ + \lambda_{\theta}\mathcal{L}_{\theta^{\text{full}}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\Delta}\mathcal{L}_{\Delta},\end{aligned} \tag{10}$$

where $\lambda_{\text{v3d}} = \lambda_{\text{j3d}} = 5.0$, $\lambda_{\Delta} = 0.01$, and all other weights are 1.0.

B.3 Evaluation Protocols

For global human motion estimation, we evaluate on EMDB-2 (Kaufmann et al., 2023), which captures monocular dynamic-camera videos, and RICH (Huang et al., 2022), which provides fixed

multi-view cameras. We measure global motion accuracy in the world coordinate system using three metrics. World-Aligned MPJPE (WA-MPJPE) aligns the predicted joints of each 100-frame segment to the ground truth using a Sim(3) transformation, while World MPJPE (W-MPJPE) estimates this alignment from only the first two frames of each segment. Both metrics are reported in millimeters. We also report the Root Translation Error (RTE, %), which measures the root translation error after SE(3) alignment over the entire sequence, normalized by the total displacement to account for varying motion magnitudes. For multi-view human pose estimation, we follow prior single-frame protocols on EgoHumans (Khirodkar et al., 2023) and EgoExo4D (Grauman et al., 2024). We report W-MPJPE† after camera-only SE(3) alignment, GA-MPJPE after jointly aligning all people with a single Sim(3) transformation, and PA-MPJPE after person-wise Procrustes alignment. All three metrics are reported in meters.

C Additional Results

C.1 Qualitative Results

In this section, we present additional qualitative results under various scenarios. As illustrated in Fig. 14, our framework reconstructs humans and the scene in diverse multi-view settings, including EgoBody (Zhang et al., 2022). As shown in Fig. 15 and Fig. 16, our framework can also handle monocular and single-image scenarios. HARMONI is capable of reconstructing both humans and the scene under both multi-view and monocular setups in a single feed-forward pass, which shows its practicability.

C.2 Failure Cases

In this section, we present failure cases where our framework fails to reconstruct humans and the scene. As shown in Fig. 17, our framework exhibits limitations when handling complex poses and close human interactions. It also struggles when the head is completely invisible in the image. Furthermore, the model faces challenges in extreme zoom-in scenarios where the head occupies most of the image. We leave these limitations for future work.



Figure 14: Qualitative results in multi-view settings.

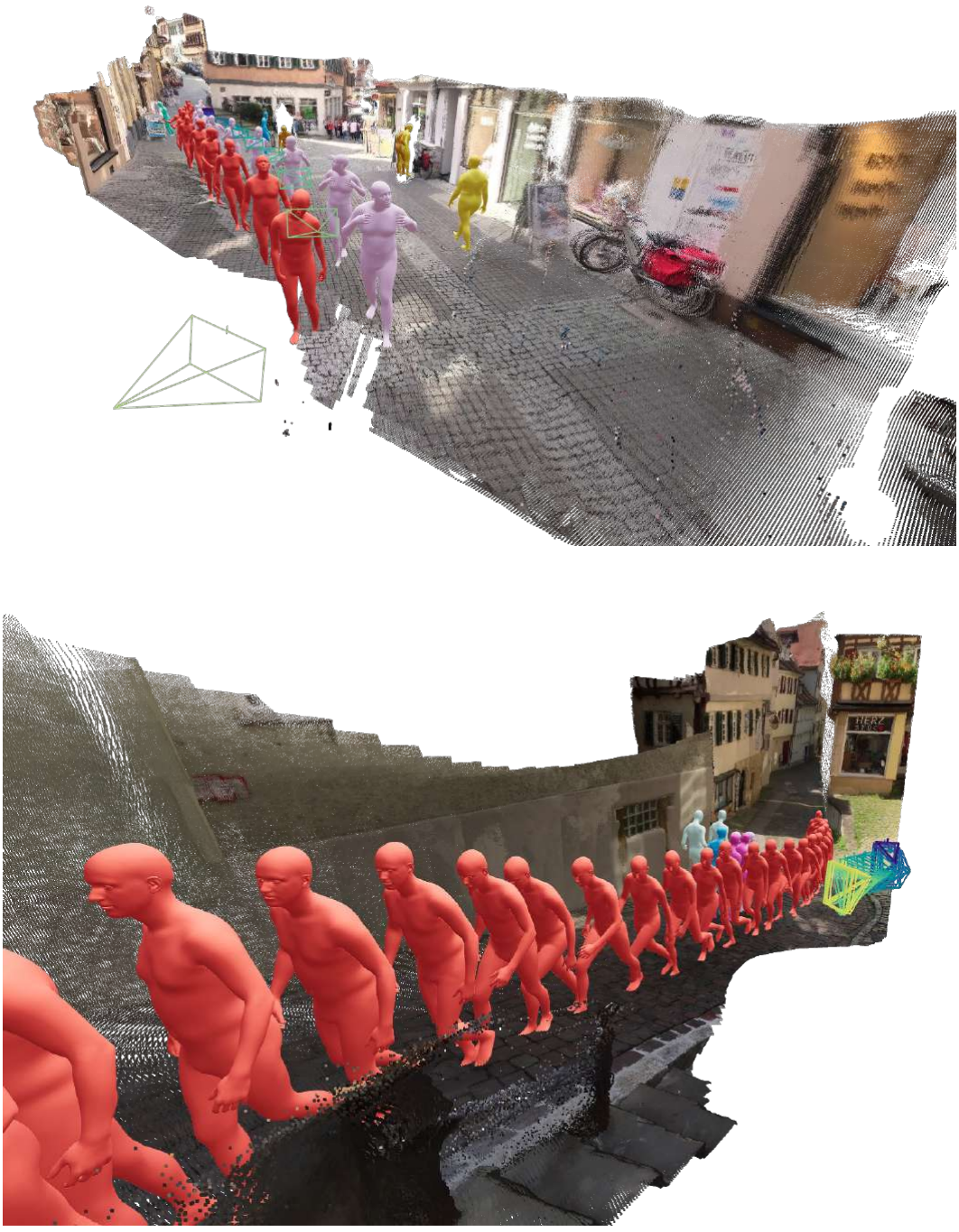

Figure 15: Qualitative results in monocular settings.

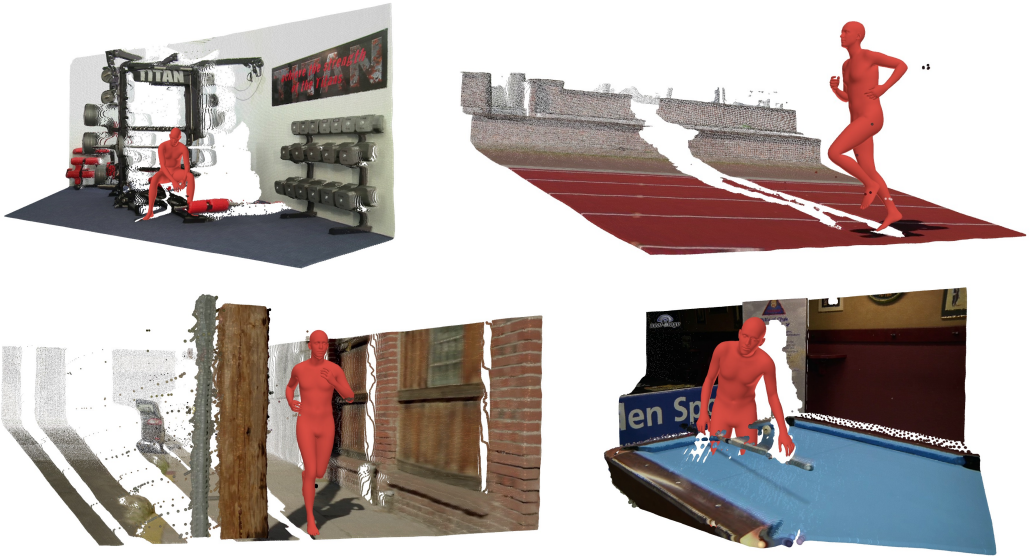

Figure 16: Qualitative results on single images.

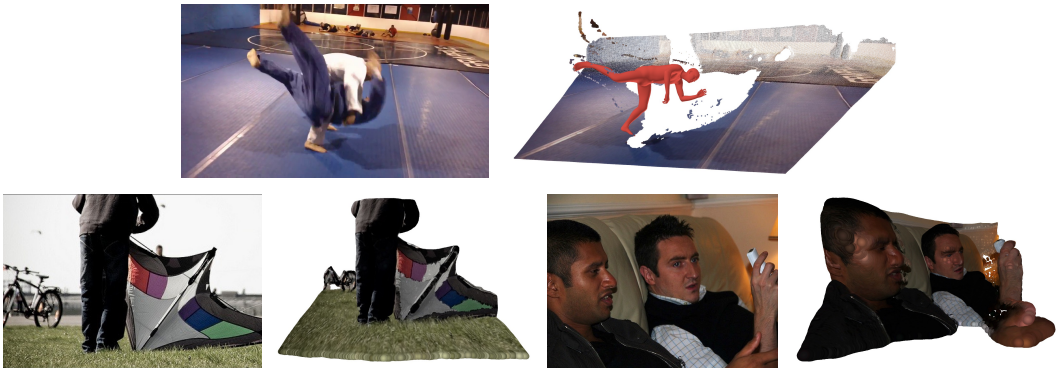

Figure 17: Failure cases.