COHERE4D: TRAINING-FREE METRIC ALIGNMENT FOR COHERENT 4D HUMAN–SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review



Figure 1: We present **COHERE4D**, a training-free framework for coherent 4D human–scene reconstruction. (a) Metric humans anchor scene scale to a unified metric world. (b) Scene geometry grounds humans, yielding better alignment and more plausible interactions. (c) Normalized performance across human–scene alignment (3DPW), scene/camera metric scale (BEDLAM2), human motion (EMDB-2) and efficiency.

ABSTRACT

We present COHERE4D, a **training-free** framework that geometrically aligns the outputs of independently pretrained human and scene foundation models to reconstruct a **coherent, metric** 4D world from monocular videos. Recent unified human–scene reconstruction methods largely learn metric scene scale and human alignment from training data, but generalize poorly to out-of-distribution scenes. Instead, COHERE4D performs test-time geometric alignment of human and scene predictions using dense pixel-aligned correspondences and RANSAC-based robust estimation. This alignment provides mutual geometric grounding: near-metric human meshes anchor the scale of scene geometry and camera translation, while scene geometry in turn refines human translation to improve global human trajectories. This test-time alignment requires no additional training and uses RANSAC-based robust estimation to suppress outliers caused by occlusion, background contamination, and erroneous depth. Across benchmarks, COHERE4D significantly outperforms prior methods in **human–scene alignment, metric-scale scene and camera pose reconstruction, and global human motion estimation**. Our training-free alignment method adds only about 4 ms/frame, while the overall inference reaches up to 10 FPS.

1 INTRODUCTION

Jointly reconstructing human, scene geometry, and camera motion from a monocular video is fundamental to 4D perception. Such a reconstruction is only physically meaningful if the human and the scene share the same metric scale. A metrically aligned human–scene representation enables geometrically consistent human–scene reasoning and supports applications ranging from AR/VR and gaming to robotic teleoperation, imitation learning, and embodied data collection (Shi et al., 2026; Lyu et al., 2026; Müller et al., 2025).

Advances in human mesh recovery and feed-forward scene reconstruction have substantially improved human and scene reconstruction quality, yet largely along independent paths. Human mesh recovery models recover metric-scale human bodies but lack reliable global scene grounding. In contrast, scene foundation models accurately estimate normalized scene geometry and camera motion, but fall short in recovering reliable metric scale.

More recent human–scene reconstruction methods have begun to address this problem through unified feed-forward architectures. Human3R (Chen et al., 2026) introduced joint human–scene recon-

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.

struction, but largely inherits CUT3R's (Wang et al., 2025b) learned metric scale without explicitly enforcing human–scene alignment. Other works explicitly address human–scene alignment through data-driven learning: UniSH (Li et al., 2026) predicts scene scale and human translation with a separate AlignNet, while SHOW (Shi et al., 2026) learns scene scale and human translation within a jointly trained human–scene model. Despite these different designs, their metric grounding still relies on learned, data-dependent scale predictions, making it vulnerable to out-of-distribution scenarios. Accurate real-world supervision for human meshes, scene geometry, camera trajectories, and metric scale is scarce, forcing large-scale training to rely on synthetic data or noisy pseudo-labels and making the learned alignment module sensitive to distribution shift. This raises a practical question: can independently pretrained human and scene foundation models be aligned into a coherent metric 4D world in a **training-free manner**?

Our key observation is that monocular scene reconstruction is inherently scale-ambiguous, whereas reconstructed parametric humans provide a strong anthropometric prior and thus a natural metric anchor. We use this human metric prior to calibrate scene scale, formulating joint human–scene reconstruction as a human-anchored metric grounding problem. Based on this insight, we introduce COHERE4D, a **training-free framework for coherent 4D human–scene reconstruction**. Given a monocular video, we first use pretrained human and scene foundation models to independently initialize human meshes, scene geometry, and camera motion. We then establish dense, pixel-aligned 3D correspondences between the visible human mesh surface and the scene prediction. Because these correspondences are inevitably contaminated by occlusion, background geometry, and erroneous depth, we introduce a robust RANSAC-based (Fischler & Bolles, 1981) alignment that separates the two quantities with different temporal properties: a sequence-level shared scale **grounds the scene and camera trajectory** to the metric human meshes, while frame-dependent **human translations correct** the dynamic placement of humans within the scene. The resulting humans, scene geometry, and cameras can therefore be transformed directly into a common metric world coordinate system, without any additional training.

Across benchmarks with diverse camera motion, occlusion patterns, and scene distributions, COHERE4D achieves more accurate human–scene geometric alignment than existing unified feed-forward approaches, while remaining robust under severe occlusion and out-of-distribution scenes. More importantly, this alignment directly corrects the learned metric scale of scene foundation models, substantially reducing metric scene-depth and camera-translation errors, which remain a key limitation of purely data-driven metric-scale prediction. Conversely, when the geometry recovered by the scene foundation model is more reliable than the global translation predicted by a human model such as Multi-HMR 2 (Fiche et al., 2026), the same alignment provides a geometric constraint that improves human global trajectories.

Our contributions are summarized as follows.

– We present COHERE4D, a simple yet effective training-free framework for coherent, metrically aligned 4D human–scene reconstruction by geometrically aligning independently pretrained human and scene foundation models at test time.

– We use metric human meshes as natural scale anchors for test-time scale calibration of scene foundation models, significantly outperforming existing 3D foundation models in metric-scale scene and camera pose reconstruction. Conversely, we demonstrate that accurate scene foundation models can help improve global human motion by grounding human trajectories in reliable scene geometry, substantially reducing full-sequence trajectory drift.

– COHERE4D achieves state-of-the-art performance in human–scene alignment (54.0% lower HSCF$_5$), metric-scale scene and camera pose reconstruction (57.1% lower scene point error), and global human motion (65.5% lower G-MPJPE). It also maintains efficient inference (5–10 FPS across backbones) and strong robustness to out-of-distribution scenes, severe occlusion, and moving cameras.

2 Related Works

2.1 Human Mesh Recovery

Human mesh recovery estimates camera-space 3D pose and shape from monocular images (Kanazawa et al., 2018; Kolotouros et al., 2019; Li et al., 2022; Goel et al., 2023; Dwivedi

et al., 2024; Patel & Black, 2025; Baradel et al., 2024; Yang et al., 2026a), with state-of-the-art methods Multi-HMR 2 (Fiche et al., 2026) and 3DB (Yang et al., 2026b) offering fast multi-person reconstruction and strong in-the-wild generalization, respectively. World-grounded methods recover global human motion using motion priors, camera motion, SLAM, gravity, and contact constraints (Yuan et al., 2022; Ye et al., 2023; Kocabas et al., 2024; Sun et al., 2023; Shin et al., 2024; Wang et al., 2024b; 2025d; 2026d;c; Zhao et al., 2026), including GVHMR's gravity-view coordinates (Shen et al., 2024) and GEM's generative motion priors (Li et al., 2025). However, these methods rarely exploit dense scene geometry to constrain trajectories, and temporal rollout can accumulate drift. We use reconstructed scene geometry to improve less reliable human trajectories and joint human–scene reconstruction.

2.2 Foundation Models for 3D/4D Reconstruction

Feed-forward 3D/4D reconstruction models predict depth or point maps, camera parameters, and scene geometry from images or videos (Wang et al., 2024a; Leroy et al., 2024; Wang et al., 2025b;a). Despite recent progress, monocular reconstruction remains inherently scale-ambiguous: models either operate in a normalized space, as in VGGT-Ω (Wang et al., 2026a), or predict metric scale using learned priors (Wang et al., 2025b; Keetha et al., 2026), including Pi3X's metric token (Wang et al., 2026b; 2025c) and DA3's metric-depth supervision (Lin et al., 2026). Absolute metric accuracy remains underexplored because common evaluation protocols emphasize relative geometry or remove scale errors through scale or Sim(3) alignment (Wang et al., 2024a; 2025a; 2026b). We instead calibrate scene geometry using more reliable human-scale priors, improving robustness under distribution shift.

2.3 Joint Human–Scene Reconstruction and Alignment

Offline optimization-based methods couple humans, cameras, and scenes: HSfM (Müller et al., 2025) uses synchronized multi-view images with human-keypoint bundle adjustment, while JOSH (Liu et al., 2026) uses human–scene contact constraints. Unlike these methods, we directly align human mesh surfaces with scene points from monocular video in an online fashion. Feed-forward methods learn joint reconstruction: Human3R (Chen et al., 2026) extends CUT3R's metric 4D prior (Wang et al., 2025b), UniSH (Li et al., 2026) predicts scene scale and human translations with AlignNet, and SHOW (Shi et al., 2026) uses scale, placement, and geometric alignment losses. Their metric grounding and human placement depend on joint training, which can limit generalization under distribution shift. More recent methods incorporate human–scene constraints through model-specific fine-tuning, using either human geometry to improve scene reconstruction (Kim et al., 2026; Xiong et al., 2025; Song et al., 2026) or scene geometry to refine humans (Sur et al., 2026; Pradyumna et al., 2026). These approaches remain tied to their fine-tuned human or scene models; in contrast, we align independently pretrained models without additional training, anchoring scene scale to metric human meshes while simultaneously using scene geometry to refine human translations. Table 1 compares representative methods.

Table 1: Comparison of representative human–scene reconstruction methods.

Method	Mono.	4D	Multi-Person	Training-Free	Online
HSfM (Müller et al., 2025)	✗	✗	✓	✓	✗
JOSH (Liu et al., 2026)	✓	✓	✓	✓	✗
MetricHMSR (Song et al., 2026)	✓	✗	✗	✗	✓
Human3R (Chen et al., 2026)	✓	✓	✓	✗	✓
UniSH (Li et al., 2026)	✓	✓	✗	✗	✓
CHROMM (Kim et al., 2026)	✓	✓	✓	✗	✓
SHOW (Shi et al., 2026)	✓	✓	✓	✗	✓
Ours	✓	✓	✓	✓	✓

3 Methods

Given a monocular RGB video $\mathcal{I} = \{\mathbf{I}_t\}_{t=1}^{T}$, where $\mathbf{I}_t \in \mathbb{R}^{H_t \times W_t \times 3}$, our goal is to reconstruct humans, scene geometry, and camera motion in a unified metric world coordinate system. As illustrated in Figure 2, we employ two independently pretrained foundation models to initialize the human and scene components. At each frame, the human branch predicts one or more metric human meshes, while the scene branch estimates dense scene geometry and camera motion (Section 3.1). Since the absolute scale predicted by scene foundation models can be inaccurate, we treat the metric human meshes as anchors and use RANSAC over multi-frame human–scene correspondences to jointly estimate a sequence-level scale and per-frame human translations; the scalar scale is used to calibrate the scene prediction to the human's metric scale (Section 3.2), while the per-frame trans-

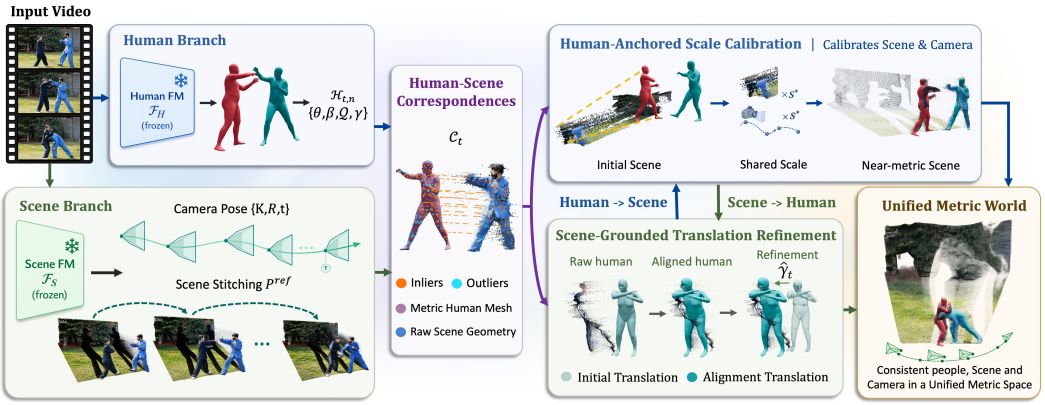

Figure 2: **Overview of COHERE4D.** Given monocular video frames, frozen human and scene branches independently reconstruct human meshes and scene geometry. Dense human–scene correspondences are established, and RANSAC selects robust inliers for geometric alignment. Human-anchored scale calibration uses humans as natural metric anchors to calibrate scene scale, while scene-grounded translation refinement uses the reconstructed scene to refine human translation, yielding a unified metric 4D world.

lations ground the humans in the calibrated scene (Section 3.3). The calibrated humans, scene geometry, and camera trajectory are finally transformed into the same metric world coordinate system. Both foundation models' weights remain unchanged throughout the pipeline, making our approach training-free.

3.1 Preliminaries

Parametric Human Models. For person n at frame t, a human reconstruction backbone $\mathcal{F}_H$ predicts pose, shape, orientation, and camera-space translation:

$$\mathcal{F}_H(\mathcal{I}) = \{\boldsymbol{\theta}_{t,n}, \boldsymbol{\beta}_{t,n}, \mathbf{Q}_{t,n}, \boldsymbol{\gamma}_{t,n}\}_{t,n}, \qquad n = 1, \ldots, N_t, \tag{1}$$

where $\boldsymbol{\theta}_{t,n}$ and $\boldsymbol{\beta}_{t,n}$ are model-specific pose and shape parameters, $\mathbf{Q}_{t,n} \in \mathrm{SO}(3)$ and $\boldsymbol{\gamma}_{t,n} \in \mathbb{R}^3$ are the global body orientation and translation in camera coordinates.

The body decoder $\mathcal{B}$ produces translation-free and camera-space metric meshes:

$$\bar{\mathbf{M}}_{t,n} = \mathcal{B}(\boldsymbol{\theta}_{t,n}, \boldsymbol{\beta}_{t,n}, \mathbf{Q}_{t,n}) \in \mathbb{R}^{N_v \times 3}, \qquad \mathbf{M}^c_{t,n} = \bar{\mathbf{M}}_{t,n} + \mathbf{1}_{N_v} \boldsymbol{\gamma}^\top_{t,n}, \tag{2}$$

where $\bar{\mathbf{M}}_{t,n}$ and $\mathbf{M}^c_{t,n}$ are the translation-free and camera-space meshes, respectively, and N_v is the model-specific vertex count. Our framework is agnostic to the underlying parametric body representation and supports diverse human reconstruction backbones (Fiche et al., 2026; Goel et al., 2023; Yang et al., 2026b; Chen et al., 2026), requiring only metric human surfaces and camera-space root transformations, regardless of parameterization or mesh topology.

Scene Foundation Models. Given a video or temporal subset, a scene foundation model $\mathcal{F}_S$ predicts camera parameters and dense geometry:

$$\mathcal{F}_S(\mathcal{I}) = \{\mathbf{K}_t, \mathbf{R}_t, \mathbf{t}_t, \mathcal{G}_t\}_t, \tag{3}$$

where $\mathbf{K}_t \in \mathbb{R}^{3\times3}$ contains camera intrinsics, $\mathcal{G}_t$ denotes dense geometry, and $(\mathbf{R}_t, \mathbf{t}_t) \in \mathrm{SO}(3) \times \mathbb{R}^3$ maps camera coordinates to the first frame of the current scene-model input.

We convert scene geometry (Wang et al., 2026a;b; Lin et al., 2026) to camera-space points by back-projecting depth maps $\mathcal{G}_t = \mathbf{D}_t \in \mathbb{R}^{H_t \times W_t}$:

$$\mathbf{p}^c_t(\mathbf{u}) = \mathbf{D}_t(\mathbf{u})\, \mathbf{K}_t^{-1} \tilde{\mathbf{u}}, \qquad \tilde{\mathbf{u}} = [u, v, 1]^\top. \tag{4}$$

For models that predict point maps, we transform them from the scene-model's reference frame to camera-space points $\mathbf{p}^c_t(\mathbf{u})$ using the predicted camera extrinsics.

3.2 Unified Human–Scene Alignment

Although some scene foundation models can predict metric-scale geometry, their data-driven scale estimates remain sensitive to distribution shift. Using reconstructed humans as near-metric references, we estimate a shared scene scale and per-frame human translations from pixel-aligned human–scene correspondences.

Overlap-Aware Scene Reconstruction. Scene foundation models typically process only a limited number of frames jointly. We therefore partition the input video into a sequence of overlapping chunks $\{\mathcal{I}^{(k)}\}_{k=1}^{K}$ along the temporal dimension, where each chunk contains at most L frames and two consecutive chunks share O overlapping frames. All frames within a chunk are jointly processed by the scene foundation model, yielding chunk-local scene geometry and camera motion.

To align chunk k with the previously reconstructed sequence, we use the shared frames to establish pixel-aligned 3D correspondences $\{(\mathbf{p}_i^{\mathrm{ref}}, \mathbf{p}_i^{\mathrm{cur}})\}_{i=1}^{N}$, where $\mathbf{p}_i^{\mathrm{ref}}$ and $\mathbf{p}_i^{\mathrm{cur}}$ denote the same overlapping pixel observed in the aligned reference reconstruction and the current chunk, respectively. We estimate a similarity transformation

$$\mathbf{p}_i^{\mathrm{ref}} \simeq \alpha_k \mathbf{A}_k \mathbf{p}_i^{\mathrm{cur}} + \mathbf{d}_k, \qquad \mathbf{A}_k \in \mathrm{SO}(3), \tag{5}$$

where $\alpha_k \in \mathbb{R}_+$, $\mathbf{A}_k$, and $\mathbf{d}_k \in \mathbb{R}^3$ denote the inter-chunk scale, rotation, and translation, respectively. We first estimate α_k from overlap depth ratios using a weighted median with outlier rejection based on the median absolute deviation (MAD). With α_k fixed, we recover $(\mathbf{A}_k, \mathbf{d}_k)$ using confidence-weighted Procrustes within iteratively reweighted least squares (IRLS) (Holland & Welsch, 1977), which downweights high-residual correspondences. The estimated transformation is applied to both scene geometry and camera motion. By sequentially composing these transformations, we stitch all scene chunks and camera poses into the first chunk's coordinate system, yielding a temporally consistent long-sequence reconstruction in a chunk-wise online manner.

Dense Human–Scene Surface Correspondences. Simultaneously with scene reconstruction, we estimate a sequence-level metric scene scale from the first chunk using the metric human surface as a geometric reference. For clarity, we omit the person index n below and describe one human at frame t. For multi-person inputs, we construct correspondences separately for each valid human mesh and combine their scale evidence to estimate one shared scene scale, while estimating a separate translation for each person in each frame. We use scene-predicted intrinsics to decode human translation. We render $\mathbf{M}_t^c$ by projecting its triangles onto the image plane via $\Pi_{\mathbf{K}_t}$ and rasterizing them with a z-buffer. Let Ω_t^H denote the resulting visible human region under self-occlusion.

For each pixel $\mathbf{u}_i = (u_i, v_i) \in \Omega_t^H$, perspective-correct barycentric interpolation on the front-most triangle recovers a continuous 3D human surface point $\mathbf{m}_{t,i}$ on the surface of $\bar{\mathbf{M}}_t$ before global translation. At the same pixel, we recover the corresponding camera-space scene point $\mathbf{p}_{t,i}^c$ by back-projecting the predicted depth (Eq. 4). As shown in Figure 2, this gives the pixel-aligned 3D correspondence set

$$\mathcal{C}_t = \left\{ (\mathbf{p}_{t,i}^c, \mathbf{m}_{t,i}) \;\middle|\; \mathbf{u}_i \in \Omega_t^H,\ D_t(\mathbf{u}_i) > 0 \right\}. \tag{6}$$

Unlike nearest-neighbor matching in 3D, these correspondences are induced directly by the shared image pixel and preserve the visible human-surface association.

Joint Scale and Translation Estimation. We formulate human–scene alignment as a joint least-squares problem over a shared scale s and per-frame human translations $\{\boldsymbol{\gamma}_t\}$:

$$\min_{s>0,\{\boldsymbol{\gamma}_t\}} \sum_{t,i:\ (\mathbf{p}_{t,i}^c, \mathbf{m}_{t,i}) \in \mathcal{C}_t} \left\| s\mathbf{p}_{t,i}^c - \mathbf{m}_{t,i} - \boldsymbol{\gamma}_t \right\|_2^2, \tag{7}$$

where s converts the scene prediction to the metric human scale and $\boldsymbol{\gamma}_t$ places the translation-free human surface in the camera coordinate system. To robustly estimate these quantities, we generate per-frame scale hypotheses, select a shared scale through cross-frame consensus, and refine human translations with the selected scale fixed. For a single-frame subset $\mathcal{S} \subseteq \mathcal{C}_t$, the least-squares scale–translation hypothesis is

$$s = \frac{\sum_{(\mathbf{p}_{t,i}^c, \mathbf{m}_{t,i}) \in \mathcal{S}} (\mathbf{p}_{t,i}^c - \bar{\mathbf{p}}^c)^\top (\mathbf{m}_{t,i} - \bar{\mathbf{m}})}{\sum_{(\mathbf{p}_{t,i}^c, \mathbf{m}_{t,i}) \in \mathcal{S}} \|\mathbf{p}_{t,i}^c - \bar{\mathbf{p}}^c\|_2^2}, \qquad \boldsymbol{\gamma} = s\bar{\mathbf{p}}^c - \bar{\mathbf{m}}, \tag{8}$$

where $\bar{\mathbf{p}}^c$ and $\bar{\mathbf{m}}$ are the respective means over $\mathcal{S}$.

Background geometry, occlusion boundaries, other people, and erroneous depths can contaminate the projected human region, biasing scale and translation estimates and thus human placement. We therefore generate RANSAC hypotheses from minimal correspondence subsets and score them by inlier support using $r_{t,i}(s, \boldsymbol{\gamma}) = \|s\mathbf{p}_{t,i}^c - \mathbf{m}_{t,i} - \boldsymbol{\gamma}\|_2 \le \tau_t$, where τ_t adapts to human body size. We refine the best-supported hypothesis on its inliers with Huber weighting (Appendix A.1).

Cross-Frame Shared Metric Scale. To suppress depth-induced frame-wise scale fluctuations, we select $s^\star = s_t$ with the greatest cross-frame inlier support among valid frames in the first chunk. We apply this shared scale $s^\star$ to scene geometry and camera translation while preserving the metric human body:

$$D_t^{\text{metric}} = s^\star D_t, \qquad \mathbf{p}_t^{c,\text{metric}} = s^\star \mathbf{p}_t^c, \qquad \mathbf{t}_t^{\text{metric}} = s^\star \mathbf{t}_t. \tag{9}$$

3.3 Scene-Grounded Translation Refinement

When the reconstructed scene is more reliable than the human model's global trajectory, it provides a complementary geometric constraint for grounding the human motion in the surrounding scene. We therefore introduce scene-grounded translation refinement to refine the human translation after determining the shared metric scale $s^\star$.

For each frame, we reuse the pixel-aligned human–scene correspondences $\mathcal{C}_t$ defined in Section 3.2 and keep $s^\star$ fixed. Each correspondence directly induces a translation estimate $s^\star \mathbf{p}_{t,i}^c - \mathbf{m}_{t,i}$. We robustly aggregate these estimates using iterative inlier filtering and coordinate-wise median refinement:

$$\hat{\boldsymbol{\gamma}}_t = \text{RobustMed}_{(\mathbf{p}_{t,i}^c, \mathbf{m}_{t,i}) \in \mathcal{C}_t} \left(s^\star \mathbf{p}_{t,i}^c - \mathbf{m}_{t,i} \right), \qquad \hat{\mathbf{M}}_t^c = \bar{\mathbf{M}}_t + \mathbf{1}_{N_v} \hat{\boldsymbol{\gamma}}_t^\top. \tag{10}$$

Here, RobustMed denotes the robust translation refitting procedure detailed in Appendix A.2. Only the global human translation is updated to align the human with the reconstructed scene, while the predicted body pose, shape, and global orientation remain unchanged.

This refinement particularly benefits human models with less reliable global trajectories, such as Multi-HMR 2 (Fiche et al., 2026): it improves human–scene alignment and, when the scene geometry is the more reliable spatial reference, also improves global human motion.

4 Experiments

We evaluate COHERE4D from three complementary perspectives: human–scene geometric alignment (Section 4.1), metric scene and camera reconstruction (Section 4.2), and global human motion reconstruction (Section 4.3). We further provide ablation and additional discussion in Section 4.4.

4.1 Human–Scene Geometric Alignment

We evaluate human–scene geometric alignment on 3DPW (Von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023) in Table 2. Following SHOW (Shi et al., 2026), we use HS-CF_5 and HS-CF_{10}, which measure the one-way nearest-neighbor distance from the visible human surface to scene points within the human mask, normalized by the mean absolute depth deviation of the human surface. The human mask is generated by the same SAM3 (Carion et al., 2026) model for all methods. The subscripts 5 and 10 indicate removing the outermost 5% and 10% of scene points along each coordinate axis, respectively. Lower values indicate better human–scene alignment.

Table 2: **Quantitative comparison of human–scene alignment** on 3DPW and EMDB-1 using HS-CF_5 and HS-CF_{10}. * denotes results reported in SHOW that are not directly comparable to our reproduced results. Our method uses scale calibration and translation refinement to improve human–scene geometric alignment. Best result; Second-best result.

Method	3DPW		EMDB-1	
	HS-CF_5 ↓	HS-CF_{10} ↓	HS-CF_5 ↓	HS-CF_{10} ↓
Human3R* (Chen et al., 2026)	1.300	1.320	1.130	1.310
UniSH* (Li et al., 2026)	7.650	7.520	7.100	6.990
SHOW* (Shi et al., 2026)	5.690	5.440	5.880	5.790
Human3R	1.108	1.345	0.719	0.924
UniSH	3.834	4.320	1.166	1.528
Ours				
Multi-HMR 2 + DA3	1.074	1.308	0.571	0.792
Multi-HMR 2 + Pi3X	**0.510**	**0.806**	0.532	0.749
Multi-HMR 2 + VGGT-Ω	0.546	0.853	**0.470**	**0.691**

Since SHOW has not yet released its implementation, we report three comparison blocks. The first, marked with *, contains results reported by SHOW; the second contains our reproduced baselines; and the third evaluates our method with different human and scene foundation models. Our method combines Multi-HMR 2 (Fiche et al., 2026) with DA3 (Lin et al., 2026), Pi3X (Wang et al., 2026b), or VGGT-Ω (Wang et al., 2026a) through human-anchored scale calibration (Section 3.2) and scene-grounded translation refinement (Section 3.3).

Despite learning joint human–scene reconstruction, Human3R (Chen et al., 2026) and UniSH (Li et al., 2026) exhibit visible human–scene misalignment and implausible contact on out-

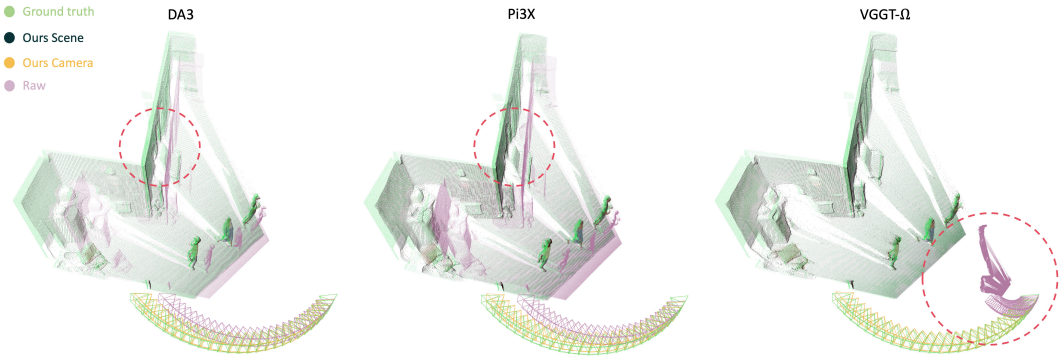

Figure 3: **Metric scene and camera motion on BEDLAM.** Our calibration brings the raw scene geometry and camera trajectory closer to the ground-truth metric scale (VGGT-Ω: normalized).

Table 3: **Metric scene reconstruction and camera pose accuracy** on Bonn and BEDLAM2. Within each scene-backbone block, Raw denotes the native scene prediction (VGGT-Ω: normalized), while our variants calibrate scene scale using Multi-HMR 2 or 3DB as metric anchors. Point Error, ATE, and RPE-t are reported in centimeters. The best result within each block is boldfaced; overall best and overall second-best.

Scene Backbone	Scale Calibration	Bonn					BEDLAM2				
		P-Err. ↓	AbsRel ↓	$\delta < 1.25$ ↑	ATE ↓	RPE-t ↓	P-Err. ↓	AbsRel ↓	$\delta < 1.25$ ↑	ATE ↓	RPE-t ↓
Human3R (Chen et al., 2026)		30.69	0.077	95.28	28.83	10.79	3692.3	0.250	57.69	26.14	11.99
UniSH (Li et al., 2026)		98.07	0.318	55.41	36.76	15.24	3506.0	0.210	66.03	70.80	35.13
DA3	Raw	29.22	0.117	93.18	6.04	3.45	3537.0	0.235	54.84	23.98	9.98
	Ours w/ Multi-HMR 2	**19.67**	**0.070**	**97.43**	**5.30**	**3.20**	3470.2	0.192	70.89	**16.29**	**6.96**
	Ours w/ 3DB	24.91	0.085	96.54	5.81	3.40	**3421.1**	**0.190**	**74.19**	17.42	7.42
Pi3X	Raw	30.48	0.090	98.49	8.70	3.53	3529.5	0.194	66.39	17.41	7.66
	Ours w/ Multi-HMR 2	**13.08**	**0.036**	98.67	6.15	3.15	**3445.7**	**0.125**	**90.01**	**10.85**	**4.97**
	Ours w/ 3DB	18.04	0.066	**98.68**	**5.65**	**3.10**	3499.8	0.169	82.45	15.96	7.01
VGGT-Ω	Raw	170.0	0.572	0.315	22.58	7.40	4404.9	0.901	0.006	111.9	44.93
	Ours w/ Multi-HMR 2	**16.75**	**0.064**	**97.72**	**3.58**	**2.16**	3288.5	**0.132**	**88.88**	**10.98**	**5.22**
	Ours w/ 3DB	28.00	0.097	97.27	3.64	2.40	**3275.5**	0.165	84.18	16.56	7.25

of-distribution inputs (Figure 4, first row). Our training-free alignment consistently achieves lower HS-CF errors on both datasets and more plausible human–scene contact. Compared with Human3R, Multi-HMR 2 + Pi3X reduces HS-CF_5 by 54.0% on 3DPW, while Multi-HMR 2 + VGGT-Ω reduces it by 34.6% on EMDB-1. Applying it to Human3R also improves consistency without backbone retraining (Appendix B.4). Stronger scene foundation models, such as Pi3X and VGGT-Ω, generally yield lower errors: their cleaner, more complete geometry provides more reliable correspondences for scale and translation estimation.

4.2 Metric Scene Reconstruction

To evaluate near-metric scene geometry and camera motion, we benchmark on Bonn (Palazzolo et al., 2019) and BEDLAM2 (Tesch et al., 2025), which provide metric ground-truth geometry and camera trajectories, as shown in Table 3. Following standard 3D reconstruction protocols (Chen et al., 2026; Wang et al., 2026b;a), we evaluate scene geometry with Point Error (P-Err.), Absolute Relative Error (AbsRel), and $\delta < 1.25$ (percentage), and camera motion with Absolute Trajectory Error (ATE) and translational Relative Pose Error (RPE-t), with all baselines re-evaluated consistently. We align the reconstructed scene and camera trajectories to the ground-truth coordinate frame using a single global rotation and translation determined from the first camera pose. We apply no ground-truth-based scale correction, preserving each method's predicted scale throughout evaluation. BEDLAM2 has a much larger scene depth range than Bonn, making its absolute 3D point errors naturally larger for comparable relative errors.

We compare against both holistic human–scene reconstruction methods and scene foundation models.

Holistic human–scene reconstruction methods. Although Human3R (Chen et al., 2026) and UniSH (Li et al., 2026) incorporate human cues, their metric scene accuracy can still be inferior to that of dedicated scene foundation models. These results highlight the difficulty of generalizing learned metric-scale prediction to out-of-distribution scenes, even when human cues are incorporated.

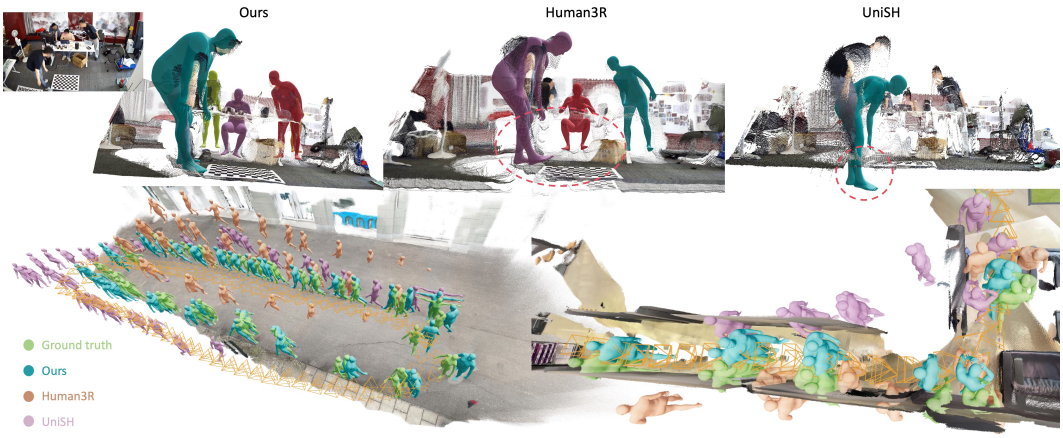

Figure 4: **Qualitative comparison of human–scene reconstruction under occlusion and global human motion** on Internet videos and EMDB-2. Top: COHERE4D yields more accurate human–scene alignment and contact than Human3R and UniSH. Bottom: COHERE4D exhibits less long-term global trajectory drift and better matches the ground truth than the baselines.

Scene foundation models. For DA3 (Lin et al., 2026) and Pi3X (Wang et al., 2026b), which natively predict metric-scale geometry, our human-anchored calibration consistently improves their raw metric predictions. On Bonn, calibrating Pi3X reduces P-Err. by up to 57.1% and ATE by 35.1%; on BEDLAM2, ATE is reduced by up to 37.7%. Figure 3 further shows that our calibrated scene and camera trajectory are closer to the ground-truth metric scale than the raw predictions. VGGT-Ω (Wang et al., 2026a), which outputs normalized, scale-ambiguous geometry, can also be calibrated into a near-metric scene and camera trajectory for world-space human reconstruction. Overall, our method substantially improves the metric-scale accuracy of scene foundation models.

4.3 Global Human Motion Reconstruction

We evaluate global human motion in world coordinates on BEHAVE (Bhatnagar et al., 2022), RICH (Huang et al., 2022), and EMDB-2 (Kaufmann et al., 2023), which respectively provide human–object interactions with frequent occlusions, challenging in-the-wild scenes, and moving-camera sequences (Table 4). Following prior protocols (Shin et al., 2024; Shen et al., 2024), we report W-MPJPE, WA-MPJPE, and Root Translation Error (RTE, %). W-MPJPE and WA-MPJPE align each 100-frame segment to ground truth using its first two frames and all frames, respectively. Following recent long-sequence evaluation (Qian et al., 2026), we additionally report Global MPJPE (G-MPJPE), which aligns the full sequence using only its first frame, exposing long-term drift.

Table 4 compares our framework across backbone combinations against existing methods. For Multi-HMR 2 (Fiche et al., 2026), scene-grounded translation refinement updates the human translation frame by frame, whereas 3DB (Yang et al., 2026b;a) retains its original camera-frame translation and combines it with the scene camera trajectory after scale calibration to obtain the world-space human trajectory. On BEHAVE (Bhatnagar et al., 2022), both Multi-HMR 2+Pi3X† and 3DB+Pi3X‡ outperform the baselines, showing that our scene grounding remains robust under object occlusion. On RICH, GEM (Li et al., 2025) remains stronger than the Multi-HMR 2 variants on W-MPJPE and WA-MPJPE, while the 3DB variants achieve better results. On the moving-camera EMDB-2 benchmark, the optimization-based JOSH (Liu et al., 2026) remains competitive with the Multi-HMR 2 variants, while our 3DB+VGGT-Ω configuration outperforms JOSH across all reported metrics without costly offline optimization. Our geometric alignment adds only about 10 ms/frame during initialization and 4 ms/frame in steady-state inference on an NVIDIA RTX 4090, while inference reaches 5–10 FPS across backbone configurations; see Appendix B.8 for details.

Most importantly, across two human backbones and three scene backbones, all of our configurations consistently achieve substantially **lower G-MPJPE across all three datasets**, indicating less full-sequence trajectory drift. On EMDB-2, our best Multi-HMR 2+DA3 and 3DB+VGGT-Ω configurations reduce G-MPJPE by 68.4% and 75.3%, respectively, relative to Human3R (Chen et al., 2026), the strongest holistic human–scene baseline with reported G-MPJPE. Compared with GVHMR (Shen et al., 2024), the best-performing baseline on EMDB-2 for this metric, Multi-HMR 2+DA3 and 3DB+VGGT-Ω reduce G-MPJPE by 55.8% and 65.5%, respectively. Both gains

Table 4: **Global human motion reconstruction in world coordinates** on BEHAVE, RICH, and EMDB-2. The first block contains human-only motion methods without explicit scene reconstruction, while the second block contains optimization-based human–scene methods, and the third block contains feed-forward human–scene reconstruction methods. † applies both scale calibration and translation refinement, while ‡ applies scale calibration only, retaining the original 3DB camera-frame translation. overall best and overall second-best.

Method	BEHAVE				RICH				EMDB-2				FPS ↑
	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	RTE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	RTE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	RTE ↓	
GVHMR	509.8	133.8	80.0	1.00	277.7	126.3	78.8	2.40	3487.5	276.5	111.0	2.00	**30.9**
GEM	550.6	151.7	90.6	1.10	314.8	118.6	75.3	1.90	3522.5	202.1	74.3	1.20	–
JOSH	–	–	–	–	–	184.3	102.6	3.40	–	174.7	68.9	1.30	0.8
JOSH3R	–	–	–	–	–	334.9	190.4	6.30	–	661.7	220.0	13.10	15.4
MetricHMSR	–	–	–	–	–	165.5	109.6	1.90	–	199.5	72.1	1.40	–
Human3R	360.3	112.2	70.3	0.73	349.3	184.9	110.0	3.30	4866.0	267.9	112.2	2.20	4.9
UniSH	349.4	239.9	134.8	1.11	409.4	183.2	118.1	4.80	6099.7	270.1	118.5	5.80	6.2
CHROMM-mono	–	–	–	–	–	138.3	87.5	3.30	–	255.0	102.6	1.70	5.1
UniCon3R	–	–	–	–	–	129.8	81.5	1.70	–	285.6	113.7	2.30	2.4
SHOW	–	–	–	–	–	172.7	107.3	2.20	–	262.3	109.1	2.10	–
Ours													
Multi-HMR 2 + DA3†	113.2	101.6	68.7	0.32	246.1	179.4	104.9	2.97	1539.9	234.4	89.9	1.41	7.2
Multi-HMR 2 + Pi3X†	112.1	93.8	64.4	0.26	211.3	151.3	92.8	2.32	3313.8	328.2	119.7	3.29	6.8
Multi-HMR 2 + VGGT-Ω†	120.1	99.9	66.7	0.28	226.9	167.0	96.3	2.70	1580.1	210.9	78.9	1.31	10.3
3DB + DA3‡	85.3	72.3	46.8	0.26	142.9	100.6	66.9	1.51	1411.2	162.4	68.6	1.21	5.8
3DB + Pi3X‡	79.7	**64.7**	**42.5**	**0.23**	145.0	100.3	66.6	1.48	2303.4	286.3	103.9	2.67	5.7
3DB + VGGT-Ω‡	**77.6**	69.7	45.4	0.25	**138.9**	**99.3**	**66.2**	**1.47**	**1203.9**	**152.9**	**68.1**	**1.15**	7.8

require no additional training. The last row of Figure 4 further shows that our human trajectories remain consistently closer to the ground truth than the baselines over long sequences. More qualitative results are provided in Appendix B.2 and Appendix B.3.

4.4 Ablation and Discussion

Table 5 compares scale calibration and translation refinement for both human backbones with Pi3X (Wang et al., 2026b); additional backbone results and analysis appear in Appendix B.5.

Table 5: **Ablation** of scale calibration and translation refinement for global human motion. *Raw* uses the original human and camera trajectories; *+scale* calibrates the camera trajectory while retaining the original human translation; *+scale&trans* further applies scene-grounded human translation refinement.

Human + Scene	Setting	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓
RICH				
Multi-HMR 2+Pi3X	+scale	271.8	209.2	119.8
	+scale&trans	**211.3**	**151.3**	**92.8**
3DB+Pi3X	+scale	**145.0**	100.3	66.6
	+scale&trans	156.5	**98.0**	**65.5**
EMDB-2				
Multi-HMR 2+Pi3X	Raw	5877.1	411.6	157.2
	+scale	**3303.5**	329.2	122.5
	+scale&trans	3313.8	**328.2**	**119.7**
3DB+Pi3X	Raw	5359.4	377.5	144.6
	+scale	**2303.4**	**286.3**	**103.9**
	+scale&trans	2308.8	288.3	105.9

We find that on the static-camera RICH (Huang et al., 2022) dataset, scene geometry is sufficiently reliable for translation refinement to consistently improve Multi-HMR 2 (Fiche et al., 2026) (*+scale&trans*). On the moving-camera EMDB-2 (Kaufmann et al., 2023), *+scale* substantially improves over *Raw*, showing that accurate metric camera translation directly benefits world-space human motion. The additional gains from *+scale&trans* depend more strongly on scene quality under camera motion, and improve as stronger scene foundation models provide more reliable foreground geometry.

In contrast to the fast multi-person reconstruction of Multi-HMR 2 (Fiche et al., 2026), 3DB (Yang et al., 2026b;a) already provides strong camera-frame translation that can be more reliable than the scene foundation models, as 3DB is trained on a vast amount of human-centric data compared to general scene foundation models. We therefore retain 3DB's original camera-frame human translation in Table 4, while translation refinement can still improve human–scene alignment, as shown in our supplementary video. Overall, our robust geometric alignment improves metric scene and camera reconstruction, while reliable scene geometry further refines less accurate human motion and preserves already strong human predictions. This bidirectional grounding consistently improves joint human–scene reconstruction.

5 Conclusion

We presented COHERE4D, a training-free framework for coherent, metrically aligned 4D human–scene reconstruction from monocular video. Our framework uses metric human reconstructions to calibrate scene scale and scene geometry to refine human translation. Experiments demonstrate gains in human–scene alignment, metric scene reconstruction, and global human motion across diverse settings, including out-of-distribution inputs and severe occlusions. Limitations and failure cases can be found in the Appendix B.11.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, proofreading, and limited code generation. The research ideas, methodological design, experimental setup, and overall technical structure of the paper were developed by the authors.

For writing, the manuscript was fully developed by the authors, with generative AI used only for language polishing and readability improvement. All AI-assisted text was subsequently reviewed, corrected, and finalized by at least two authors. For coding, generative AI was used to assist with implementing limited parts of the code based on algorithms and framework designs specified by the authors. All AI-assisted code was manually inspected and tested for correctness by the authors.

We did not use generative AI to propose research ideas or hypotheses, design the methodology or experiments, or generate scientific figures or images. We take full responsibility for the final content of this work, including all text, claims, code, experiments, and results produced with the assistance of generative AI.

ETHICS STATEMENT

This study follows the ICLR Code of Ethics. We use only publicly available research datasets and publicly accessible Internet videos, and do not collect new human-subject data. Our work focuses on geometric human–scene reconstruction and does not involve identity recognition, sensitive-attribute inference, or other forms of personal profiling. All data are used solely for research evaluation and qualitative analysis.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of our method and evaluation setup. Section 3 and Appendix A describe the method and implementation details. Appendix B provides the experimental settings, evaluation protocols, ablations, and runtime analysis used for reproducing our results.

REFERENCES

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.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15914–15925. IEEE, 2022.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

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

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1323–1333. IEEE, 2024.

Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-HMR 2: Multi-person camera-centric human detection, mesh recovery and tracking. *arXiv preprint arXiv:2606.14841*, 2026.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

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

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2282–2292, 2019.

Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977.

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.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

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

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3d human pose and shape in the wild. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14632–14643. IEEE, 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.

Sangmin Kim, Minhyuk Hwang, Geonho Cha, Dongyoon Wee, and Jaesik Park. Coherent human-scene reconstruction from multi-person multi-view video in a single pass. *arXiv preprint arXiv:2603.12789*, 2026.

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

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2252–2261, 2019.

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.

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, 2025.

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

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pp. 590–606. Springer, 2022.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 141261–141285, 2026.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. In *International Conference on Learning Representations*, volume 2026, pp. 155959–155977, 2026.

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. Reviv: Reconstructing the viewer and the view in 4d from monocular egocentric video. *arXiv preprint arXiv:2607.17790*, 2026.

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21948–21958. IEEE, 2025.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7855–7862. IEEE, 2019.

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

Y. M. Pradyumna, Yuxuan Xue, Yue Chen, Nikita Kister, István Sárándi, and Gerard Pons-Moll. GRAFT: Geometric refinement and fitting transformer for human scene reconstruction. In *Computer Vision – ECCV 2026*, volume 17002 of *Lecture Notes in Computer Science*, pp. 250–267. Springer, 2026. doi: 10.1007/978-3-032-36969-7_14.

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

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

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

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

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

Tanuj Sur, Shashank Tripathi, Nikos Athanasiou, Ha Linh Nguyen, Kai Xu, Michael J Black, and Angela Yao. Unicon3r: Unified contact-aware 4d human-scene reconstruction from monocular video. *arXiv preprint arXiv:2604.19923*, 2026.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European conference on computer vision*, pp. 601–617. Springer, 2018.

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

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

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.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3X: Engineering update to π^3. `https://github.com/yyfz/Pi3`, 2025c. Official repository, Pi3X release.

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*, volume 2026, pp. 10481–10497, 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.

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

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

Zhumei Wang, Zechen Hu, Ruoxi Guo, Huaijin Pi, Ziyong Feng, Liang Zhang, Mingtao Pei, and Siyuan Huang. Mocap-2-to-3: Multi-view lifting for monocular motion recovery with 2d pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42869–42878, June 2026d.

Weitao Xiong, Zhiyuan Yuan, Jiahao Lu, Chengfeng Zhao, Peng Li, and Yuan Liu. Huprior3r: Incorporating human priors for better 3d dynamic reconstruction from monocular videos. *arXiv preprint arXiv:2512.06368*, 2025.

Timing Yang, Sicheng He, Hongyi Jing, Jiawei Yang, Zhijian Liu, Chuhang Zou, and Yue Wang. Fast SAM 3d body: Accelerating SAM 3d body for real-time full-body human mesh recovery. In *Computer Vision – ECCV 2026*, volume 17012 of *Lecture Notes in Computer Science*, pp. 262–279. Springer, 2026a. doi: 10.1007/978-3-032-37044-0_15.

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

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

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17016–17027. IEEE, 2023.

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

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. *arXiv preprint arXiv:2603.17355*, 2026.

Appendix

Table of Contents

A Method **16**

A.1 Unified Human–Scene Alignment 16

A.2 Scene-Grounded Translation Refinement 16

B Experiments **17**

B.1 Metric Scene Reconstruction on BEDLAM 17

B.2 Additional Human Trajectory Results 18

B.3 Qualitative Comparison with Baselines 20

B.4 Improving Human–Scene Alignment in Existing Baselines 21

B.5 Ablation of Scale Calibration and Translation Refinement 22

B.6 Ablation of Robust Metric Calibration 23

B.7 Qualitative Ablation of Scale Calibration and Translation Refinement 23

B.8 Runtime Evaluation 24

B.9 Out-of-Distribution Qualitative Results 24

B.10 Qualitative Results on In-the-Wild Internet Videos 25

B.11 Limitations and Failure Cases 25

A METHOD

A.1 UNIFIED HUMAN–SCENE ALIGNMENT

Human–scene correspondence construction. When more than 1,024 valid human pixels are available, we sample 1,024 pixels without replacement using a frame-dependent deterministic random seed. At each selected pixel, the visible human surface point is recovered using z-buffering and perspective-correct barycentric interpolation.

RANSAC configuration. Since human and scene points are expressed in the same camera coordinate frame, each minimal hypothesis samples two non-degenerate human–scene correspondences and solves for scale and translation in closed form. We evaluate one hypothesis fitted using all correspondences and 1,000 hypotheses fitted using randomly sampled minimal subsets. The adaptive inlier threshold is

$$\tau_t = \max\left(0.04\ \mathrm{m}, 0.02\ \mathrm{diag}(\bar{\mathbf{M}}_t)\right). \tag{11}$$

where $\mathrm{diag}(\bar{\mathbf{M}}_t)$ denotes the diagonal length of the 3D bounding box of the translation-free human mesh. Candidate hypotheses are ranked by decreasing inlier count, breaking ties by increasing median and then 90th-percentile inlier residuals. The selected hypothesis is refitted on its inliers and further refined with up to three Huber-weighted iterations (Huber, 1992), which downweight large-residual correspondences, using a transition point $\delta_t = \tau_t/2$. Our RANSAC solver parameterizes the alignment as $\mathbf{m}_{t,i} \simeq s_t \mathbf{p}^c_{t,i} + \mathbf{b}_t$. Accordingly, the human translation that aligns the translation-free body surface to the scaled scene satisfies $\mathbf{m}_{t,i} + \hat{\boldsymbol{\gamma}}_t \simeq s_t \mathbf{p}^c_{t,i}$, and is therefore given by $\hat{\boldsymbol{\gamma}}_t = -\mathbf{b}_t$. The offset $\mathbf{b}_t$ is thus an internal fitting variable rather than the final camera-space human translation.

Shared-scale selection. The sequence-level scale is estimated from the first scene chunk. We discard frame-level RANSAC scale candidates with an inlier ratio below 0.25 or an inlier median residual above 0.08 m; if none pass, we retain all finite positive candidates. For each retained candidate $s^{(k)}$, we fix the scale and independently re-estimate the translation offset in every valid frame as

$$\mathbf{b}_t^{(k)} = \underset{i}{\mathrm{median}} \left(\mathbf{m}_{t,i} - s^{(k)} \mathbf{p}^c_{t,i}\right). \tag{12}$$

The corresponding inlier sets are then refined and aggregated across the first chunk. We rank each candidate by its total cross-frame inlier support, using the median and 90th-percentile residuals as secondary criteria:

$$s^\star = \underset{s^{(k)}}{\arg\max} \left(N_{\mathrm{in}}^{(k)}, -\mathrm{Med}^{(k)}, -\mathrm{P90}^{(k)}\right), \tag{13}$$

where the tuple is compared lexicographically. The selected $s^\star$ remains fixed for the sequence, while the frame-dependent human translation is re-estimated under this fixed scale as described in Sec. 3.3.

A.2 SCENE-GROUNDED TRANSLATION REFINEMENT

After estimating the sequence-level shared scale $s^\star$, we keep it fixed for all subsequent frames and refine only the camera-space human translation. The human pose, shape, and global orientation remain unchanged.

Camera-frame Translation. Different human backbones use different camera parameterizations before producing the camera-frame human translation used in our alignment. Following Sec. 3.1, we denote the translation of person n at frame t by $\boldsymbol{\gamma}_{t,n} \in \mathbb{R}^3$. To ensure pixel-aligned human–scene correspondence, we express the camera-frame translation from all human backbones using the scene-predicted intrinsics $\mathbf{K}_t$.

For HMR2.0 (Goel et al., 2023) and 3DB (Yang et al., 2026b), the camera head predicts crop-based camera parameters $(s^{\mathrm{cam}}, t_x, t_y)$, which we convert to a full-perspective translation using the detected bounding box and scene-predicted intrinsics $\mathbf{K}_t$. Given the bounding-box center (c_x, c_y), box size b, focal length f, and principal point (c_x^K, c_y^K),

$$T_z = \frac{2f}{s^{\mathrm{cam}} b}, \qquad T_x = t_x + \frac{2(c_x - c_x^K)}{s^{\mathrm{cam}} b}, \qquad T_y = t_y + \frac{2(c_y - c_y^K)}{s^{\mathrm{cam}} b}, \tag{14}$$

and $\boldsymbol{\gamma}_{t,n} = (T_x, T_y, T_z)^\top$.

In contrast, Multi-HMR 2 (Fiche et al., 2026) directly predicts a full-perspective pelvis translation in camera coordinates. We preserve its predicted pelvis depth $T_{z,t,n}$ and image location $\tilde{\mathbf{u}}_{t,n}$, but re-express the translation using the scene-predicted intrinsics $\mathbf{K}_t$:

$$\boldsymbol{\gamma}_{t,n} = T_{z,t,n}\, \mathbf{K}_t^{-1} \tilde{\mathbf{u}}_{t,n}, \tag{15}$$

where $\tilde{\mathbf{u}}_{t,n}$ is the homogeneous pelvis image coordinate. Thus, all human translations used by our alignment are defined under the same camera intrinsics $\mathbf{K}_t$ as the scene branch.

For all human backbones, $\boldsymbol{\gamma}_{t,n}$ henceforth denotes the camera-frame human translation expressed under the scene-predicted intrinsics $\mathbf{K}_t$. For notational simplicity, we omit n below and write $\boldsymbol{\gamma}_t$, while the final refined translation is $\hat{\boldsymbol{\gamma}}_t$.

Per-frame correspondence construction. For each frame t, we use the original camera-frame human translation to project the human mesh onto the scene image plane and construct the same pixel-aligned 3D correspondences as in Sec. 3.2. Specifically, rasterization and z-buffering determine the visible human surface, while perspective-correct barycentric interpolation recovers the continuous human surface point $\mathbf{m}_{t,i}$ at each valid pixel. The corresponding scene point $\mathbf{p}^c_{t,i}$ is obtained from the scene point map at the same pixel. At most 1,024 valid correspondences are sampled.

Fixed-scale robust translation refitting. With $s^\star$ fixed, each correspondence provides a candidate human translation

$$\boldsymbol{\gamma}^{\text{cand}}_{t,i} = s^\star \mathbf{p}^c_{t,i} - \mathbf{m}_{t,i}. \tag{16}$$

We initialize the refined translation using the coordinate-wise median

$$\boldsymbol{\gamma}^{(0)}_t = \text{median}_i \left(\boldsymbol{\gamma}^{\text{cand}}_{t,i} \right).$$

At iteration k, we compute the residual $r^{(k)}_{t,i} = \|\mathbf{m}_{t,i} + \boldsymbol{\gamma}^{(k)}_t - s^\star \mathbf{p}^c_{t,i}\|_2$ and retain correspondences satisfying $r^{(k)}_{t,i} \le \tau_t$, where

$$\tau_t = \max \left(0.04 \text{ m}, 0.02 \text{ diag}(\bar{\mathbf{M}}_t) \right). \tag{17}$$

The translation is then updated by taking the coordinate-wise median of $\boldsymbol{\gamma}^{\text{cand}}_{t,i}$ over the retained inliers. We perform at most three iterations and stop early when the translation change is below 10^{-7}. If no valid refinement is obtained, we retain the original human translation $\boldsymbol{\gamma}_t$.

World-space reconstruction. After refinement, the aligned camera-space human mesh is $\hat{\mathbf{M}}^c_t = \bar{\mathbf{M}}_t + \mathbf{1}_{N_v} \hat{\boldsymbol{\gamma}}_t^\top$. Using the calibrated scene camera trajectory, the human is transformed into the common world coordinate system, where the human and scene are spatially aligned and the scene geometry and camera translation share the same metric scale $s^\star$.

B Experiments

B.1 Metric Scene Reconstruction on BEDLAM

Table 6 provides additional evaluation of metric scene geometry and camera trajectories on BEDLAM (Black et al., 2023). Since Human3R (Chen et al., 2026) and UniSH (Li et al., 2026) are trained on BEDLAM (Black et al., 2023), we exclude them for a fair comparison and report only the raw outputs of scene foundation models and their scale-calibrated outputs using Multi-HMR 2 (Fiche et al., 2026) as the metric human anchor. P-Err., ATE, and RPE-t are reported in centimeters, while AbsRel is unitless. Our calibration again yields more accurate metric scale than the native metric predictions of the scene

Table 6: Metric scene reconstruction and camera pose accuracy on BEDLAM. Raw denotes the native output of each scene foundation model. Our scale calibration uses Multi-HMR 2 as the metric human anchor. Lower is better.

Scene Backbone	Scale Calibration	P-Err. ↓	AbsRel ↓	ATE ↓	RPE-t ↓
DA3	Raw	2024.9	0.417	191.0	76.51
	Ours w/ Multi-HMR 2	**1645.1**	**0.338**	**114.6**	**47.46**
Pi3X	Raw	1909.8	0.311	228.8	93.67
	Ours w/ Multi-HMR 2	**1261.8**	**0.161**	**108.1**	**47.75**
VGGT-Ω	Raw	3155.9	0.915	576.4	228.0
	Ours w/ Multi-HMR 2	**993.4**	**0.268**	**24.83**	**10.34**

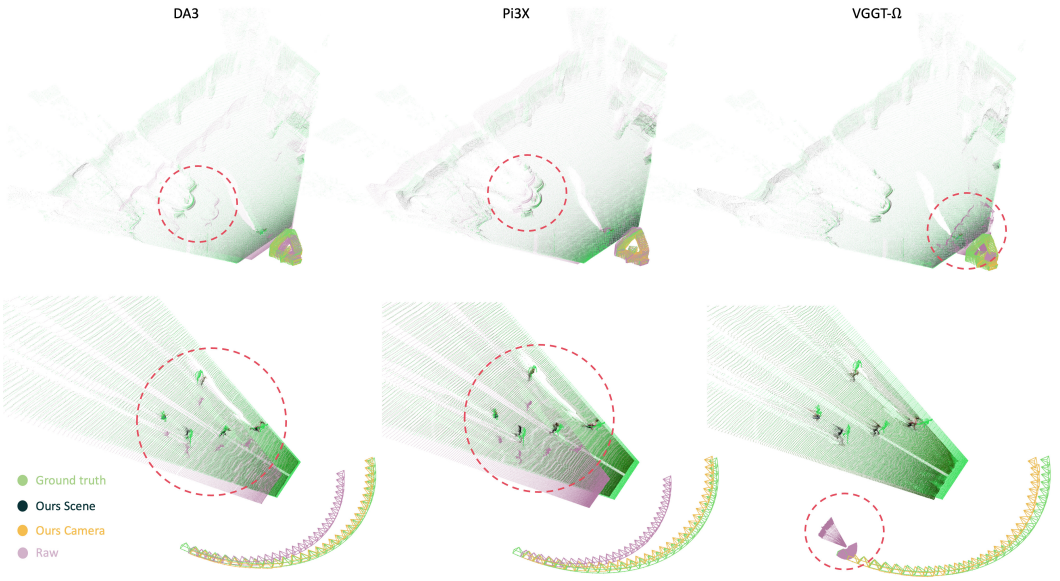

Figure 5: **Metric scene and camera motion on BEDLAM and BEDLAM2.** Our calibration brings the raw scene geometry and camera trajectory closer to the ground-truth metric scale. Specifically, VGGT-Ω produces normalized predictions, and our scale calibration converts them into near-metric reconstructions.

Table 7: Additional global human motion reconstruction results on BEHAVE, RICH, and EMDB-2. We further evaluate Multi-HMR 2 and HMR2.0 with different scene backbones, including DA3, Pi3X, and VGGT-Ω. † applies scale calibration followed by translation refinement. Best result; Second-best result. Lower is better for all metrics.

Method	BEHAVE				RICH				EMDB-2			
	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	RTE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	RTE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	RTE ↓
GVHMR	509.8	133.8	80.0	1.00	277.7	126.3	78.8	2.40	3487.5	276.5	111.0	2.00
GEM	550.6	151.7	90.6	1.10	314.8	**118.6**	**75.3**	1.90	3522.5	202.1	74.3	**1.20**
JOSH	–	–	–	–	–	184.3	102.6	3.40	–	**174.7**	**68.9**	1.30
JOSH3R	–	–	–	–	–	334.9	190.4	6.30	–	661.7	220.0	13.10
MetricHMSR	–	–	–	–	–	165.5	109.6	1.90	–	199.5	72.1	1.40
UniSH	349.4	239.9	134.8	1.11	409.4	183.2	118.1	4.80	6099.7	270.1	118.5	5.80
CHROMM-mono	–	–	–	–	–	138.3	87.5	3.30	–	255.0	102.6	1.70
UniCon3R	–	–	–	–	–	129.8	81.5	**1.70**	–	285.6	113.7	2.30
SHOW	–	–	–	–	–	172.7	107.3	2.20	–	262.3	109.1	2.10
Ours												
Multi-HMR 2 + DA3†	113.2	101.6	68.7	0.32	246.1	179.4	104.9	2.97	**1539.9**	234.4	89.9	1.41
Multi-HMR 2 + Pi3X†	112.1	93.8	64.4	**0.26**	**211.3**	151.3	92.8	2.32	3313.8	328.2	119.7	3.29
Multi-HMR 2 + VGGT-Ω†	120.1	99.9	66.7	0.28	226.9	167.0	96.3	2.70	1580.1	210.9	78.9	1.31
HMR2.0 + DA3†	117.9	105.5	65.3	0.38	270.1	159.7	97.5	2.78	2293.8	300.2	112.2	1.69
HMR2.0 + Pi3X†	109.3	**93.4**	**58.3**	0.29	234.5	130.5	83.3	1.95	2869.3	372.5	131.3	2.96
HMR2.0 + VGGT-Ω†	**105.2**	95.8	59.1	0.28	252.5	143.1	85.9	2.49	1825.2	261.7	88.8	1.41

foundation models. Although VGGT-Ω (Wang et al., 2026a) outputs normalized, scale-ambiguous geometry, its stronger human-foreground reconstruction provides more reliable correspondences, leading to lower scale error after our alignment. Figure 5 qualitatively compares raw scene geometry and camera motion with the results after scale calibration. For DA3 (Lin et al., 2026) and Pi3X (Wang et al., 2026b), the raw predictions show noticeable scale and camera-trajectory deviations from the ground truth in both indoor and outdoor scenes, whereas our scale calibration brings both closer to the correct metric scale. For the normalized, scale-ambiguous VGGT-Ω (Wang et al., 2026a) predictions, our scale calibration similarly recovers near-metric scene geometry and camera motion.

B.2 Additional Human Trajectory Results

Quantitative Results. Table 7 provides additional global human motion results on BEHAVE (Bhatnagar et al., 2022), RICH (Huang et al., 2022), and EMDB-2 (Kaufmann et al., 2023).

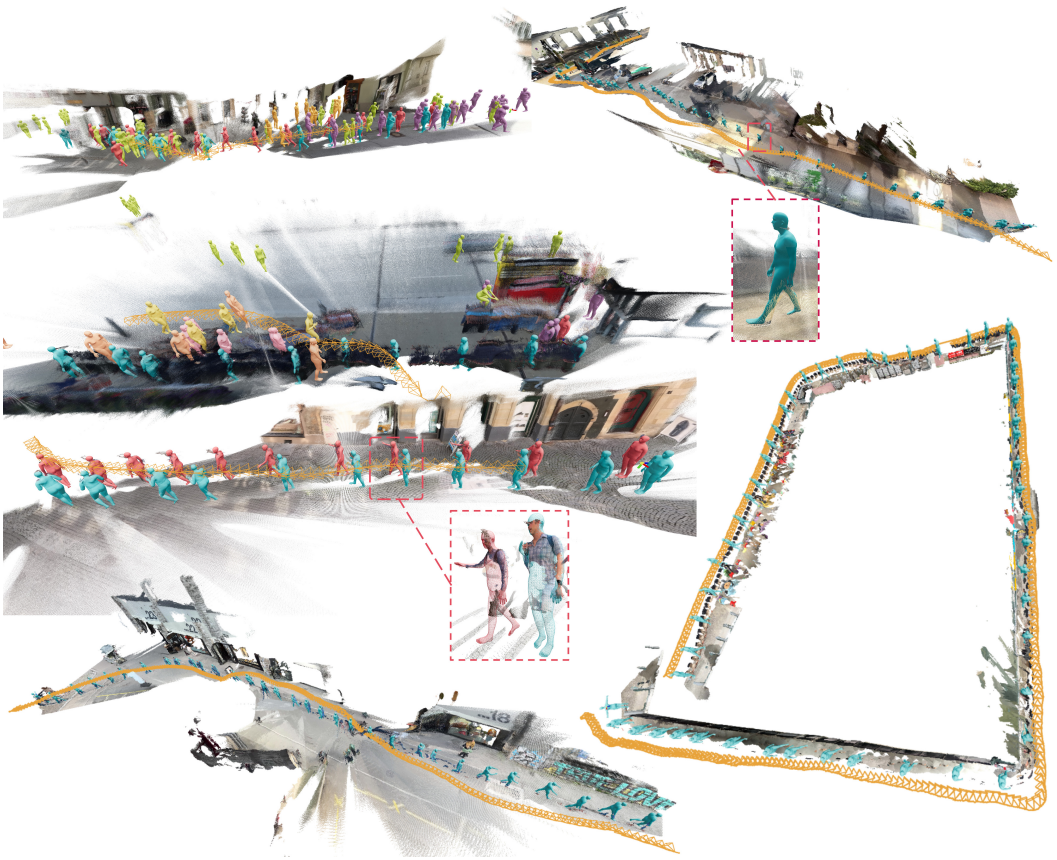

Figure 6: **Qualitative results of human–scene reconstruction** on 3DPW and EMDB-2, including multi-person scenes and long sequences.

BEHAVE (Bhatnagar et al., 2022) contains human–object interactions with frequent occlusions, making it well suited for evaluating the robustness of our alignment. For consistency, all BEHAVE evaluations use the same subset, Sub05 with camera 2, where Sub05 provides more long sequences and camera 2 is randomly selected. RICH (Huang et al., 2022) provides challenging in-the-wild scenes for testing out-of-distribution generalization, while EMDB-2 (Kaufmann et al., 2023) evaluates performance under moving-camera conditions.

We additionally evaluate HMR2.0 (Goel et al., 2023) with different scene foundation models, since several prior baselines, such as GVHMR (Shen et al., 2024), are built on this earlier human backbone. Using the same HMR2.0 backbone provides a more controlled comparison and shows that the gains of COHERE4D do not simply come from using stronger human reconstruction models. Across long-sequence evaluation, all of our variants achieve lower G-MPJPE than prior baselines, including GVHMR, with trends consistent with those observed using Multi-HMR 2 (Fiche et al., 2026). These results show that COHERE4D's long-term trajectory improvements extend to the earlier HMR2.0 backbone. Moreover, as a model-independent, training-free framework, COHERE4D can be readily combined with newer human and scene foundation models. Unlike unified baselines that require costly model retraining to incorporate new backbones, our framework can directly replace the corresponding pretrained modules, allowing it to benefit from future advances in human and scene foundation models with minimal additional cost.

Qualitative Results. Figure 6 shows qualitative human–scene reconstruction results on 3DPW (Von Marcard et al., 2018) and EMDB-2 (Kaufmann et al., 2023). Our framework reconstructs multiple people together with the scene in a shared metric space within a single inference pipeline. Over long horizons, the recovered human trajectories remain consistently aligned with the reconstructed scene.



Figure 7: **Qualitative comparison of human–scene alignment** on EMDB-2. COHERE4D yields more accurate human–scene alignment and contact than Human3R and UniSH.

B.3 Qualitative Comparison with Baselines

Human–Scene Alignment and Interaction. Figure 7 and Figure 8 qualitatively compare human–scene alignment on EMDB-2 (Kaufmann et al., 2023) and BEHAVE (Bhatnagar et al., 2022). On EMDB-2 (Kaufmann et al., 2023) (Figure 7), our improved human–scene alignment leads to more plausible object contact, e.g., the reconstructed humans are correctly supported by benches and other supporting surfaces. In contrast, inaccurate human scale in Human3R (Chen et al., 2026) can cause the body to intersect objects or be vertically misaligned with the supporting surface, while UniSH (Li et al., 2026) shows residual human–scene misalignment that shifts the contact location. Under moving cameras, our reconstructed scene also remains more consistent across frames, whereas Human3R (Chen et al., 2026) and UniSH (Li et al., 2026) may reconstruct the same object at inconsistent locations, producing visible ghosting artifacts. Similar behavior is observed on BEHAVE (Bhatnagar et al., 2022) (Figure 8), where our meshes better align with and interact with the objects. This also highlights a complementary benefit of scene-grounded translation refinement: beyond improving global trajectories when human translation is less reliable, it can improve local human–scene consistency in interaction-heavy scenes. Even when a strong baseline sometimes shows limited gains in global trajectory metrics, these metrics may not fully capture the benefits of translation refinement. In our in-the-wild videos, enabling translation refinement consistently yields visibly better human–scene alignment and more plausible human–scene contact, which is particu-

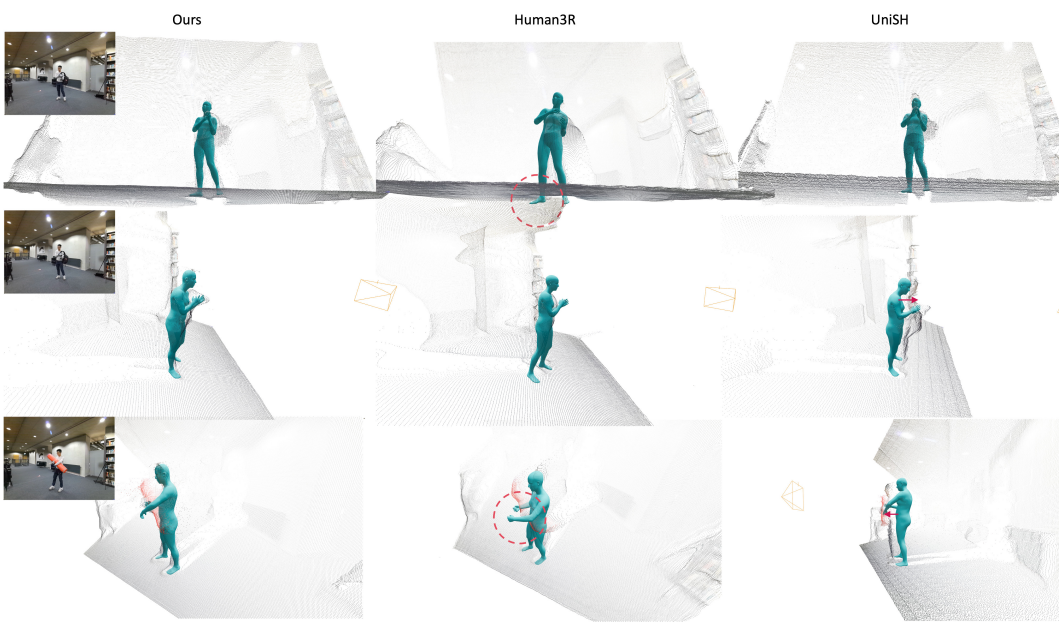

Figure 8: **Qualitative comparison of human–scene alignment** on BEHAVE. COHERE4D achieves more accurate human–scene alignment and more plausible human–object contact than Human3R and UniSH.

larly beneficial for interaction-heavy scenes. Overall, the improved human–scene alignment yields more geometrically consistent human–object interactions.

Long-Horizon Global Human Motion. Figure 9 qualitatively compares global human motion on EMDB-2 (Kaufmann et al., 2023). For all methods, the reconstructed scene and human trajectory are aligned to the ground truth only at the first frame. Over long sequences, our trajectories remain consistently closer to the ground truth with substantially less accumulated drift, consistent with the lower G-MPJPE in Table 4, while preserving better alignment with the reconstructed scene. The last row further compares the estimated camera trajectories. Our calibrated camera motion remains closer to the ground truth over time, which in turn leads to more accurate world-space human trajectories under moving cameras.

B.4 Improving Human–Scene Alignment in Existing Baselines

We further evaluate whether COHERE4D can improve an existing unified human–scene reconstruction model without retraining.

Table 8 compares the original Human3R (Chen et al., 2026) and UniSH (Li et al., 2026) predictions with their corresponding baselines w/ COHERE4D. Our human-anchored scale calibration brings the scene to the human metric scale, followed by scene-grounded translation refinement. Applying our geometric calibration consistently reduces HS-CF_5 and HS-CF_{10} for both methods on 3DPW (Von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023).

Table 8: Ablation of applying COHERE4D to Human3R and UniSH for human–scene geometric alignment on 3DPW and EMDB-1. The baselines w/ COHERE4D apply scale calibration and translation refinement to the baselines' human and scene predictions. Lower is better, and the best result within each baseline block is boldfaced.

Method	3DPW		EMDB-1	
	HS-CF_5 ↓	HS-CF_{10} ↓	HS-CF_5 ↓	HS-CF_{10} ↓
Human3R	1.108	1.345	0.719	0.924
Human3R w/ COHERE4D	**0.793**	**1.035**	**0.579**	**0.782**
UniSH	3.834	4.320	1.166	1.528
UniSH w/ COHERE4D	**0.438**	**0.618**	**0.653**	**0.958**

Figure 10 further shows qualitative ablations for Human3R and UniSH before and after applying COHERE4D. The original predictions exhibit noticeable mesh–scene misalignment, while our alignment brings the human mesh surfaces into closer agreement with the reconstructed scene. These

Table 9: **Ablation of scale calibration and translation refinement for global human motion** with Multi-HMR 2, HMR2.0, and 3DB on BEHAVE, RICH, and EMDB-2. *Raw*: original human/camera trajectories; *+scale*: camera-scale calibration; *+scale&trans*: additional scene-grounded human translation refinement. Lower is better, and the best result for each human–scene backbone combination is boldfaced.

Scene Backbone	Setting	Multi-HMR 2			HMR2.0			3DB		
		G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓
BEHAVE										
DA3	+scale	125.8	105.1	70.0	143.1	110.4	68.7	**85.3**	**72.3**	**46.8**
	+scale&trans	**113.2**	**101.6**	**68.7**	**117.9**	**105.5**	**65.3**	101.3	84.7	54.5
Pi3X	+scale	123.4	98.8	67.7	139.3	105.0	65.8	**79.7**	**64.7**	**42.5**
	+scale&trans	**112.1**	**93.8**	**64.4**	**109.3**	**93.4**	**58.3**	92.5	71.2	47.1
VGGT-Ω	+scale	124.7	103.2	69.8	130.6	104.4	65.6	**77.6**	**69.7**	**45.4**
	+scale&trans	**120.1**	**99.9**	**66.7**	**105.2**	**95.8**	**59.1**	100.7	79.1	50.2
RICH										
DA3	+scale	269.7	209.2	120.0	311.0	185.7	113.1	**142.9**	**100.6**	**66.9**
	+scale&trans	**246.1**	**179.4**	**104.9**	**270.1**	**159.7**	**97.5**	192.4	131.2	81.3
Pi3X	+scale	271.8	209.2	119.8	316.3	185.2	112.6	**145.0**	100.3	66.6
	+scale&trans	**211.3**	**151.3**	**92.8**	**234.5**	**130.5**	**83.3**	156.5	**98.0**	**65.5**
VGGT-Ω	+scale	267.3	207.4	119.4	314.8	182.6	111.4	**138.9**	**99.3**	**66.2**
	+scale&trans	**226.9**	**167.0**	**96.3**	**252.5**	**143.1**	**85.9**	184.9	118.3	70.7
EMDB-2										
DA3	Raw	2073.6	251.0	89.6	2593.9	307.1	105.3	1855.4	182.3	73.3
	+scale	**1520.1**	**216.1**	**80.6**	**2253.9**	**280.1**	**98.1**	**1411.2**	**162.4**	**68.6**
	+scale&trans	1539.9	234.4	89.9	2293.8	300.2	112.2	1452.9	189.1	84.6
Pi3X	Raw	5877.1	411.6	157.2	5493.1	446.2	168.6	5359.4	377.5	144.6
	+scale	**3303.5**	329.2	122.5	2871.4	374.9	134.4	**2303.4**	**286.3**	**103.9**
	+scale&trans	3313.8	**328.2**	**119.7**	**2869.3**	**372.5**	**131.3**	2308.8	288.3	105.9
VGGT-Ω	+scale	**1573.8**	211.2	80.8	1828.1	264.8	92.3	**1203.9**	**152.9**	**68.1**
	+scale&trans	1580.1	**210.9**	**78.9**	**1825.2**	**261.7**	**88.8**	1208.6	157.0	71.8

results demonstrate that COHERE4D can be directly plugged into existing human–scene reconstruction pipelines to improve geometric consistency.

B.5 Ablation of Scale Calibration and Translation Refinement

Table 9 provides a comprehensive ablation of Multi-HMR 2 (Fiche et al., 2026), HMR2.0 (Goel et al., 2023), and 3DB (Yang et al., 2026b) with DA3 (Lin et al., 2026), Pi3X (Wang et al., 2026b), and VGGT-Ω (Wang et al., 2026a) on BEHAVE (Bhatnagar et al., 2022), RICH (Huang et al., 2022), and EMDB-2 (Kaufmann et al., 2023). *Raw* retains the original human and camera motion; *+scale* preserves the camera-frame human translation while calibrating the scene scale and camera translation; and *+scale&trans* additionally refines the human translation using the reconstructed scene.

On the static-camera BEHAVE and RICH datasets, translation refinement consistently improves Multi-HMR 2 and HMR2.0 across all scene backbones, showing that reliable scene geometry can effectively ground their less accurate global trajectory. On the moving-camera EMDB-2 (Kaufmann et al., 2023) dataset, *+scale* substantially improves over *Raw*, demonstrating that more accurate metric camera translation directly benefits world-space human motion reconstruction. The additional gains from translation refinement are less consistent on EMDB-2 (Kaufmann et al., 2023), where camera motion can degrade the scene foundation model's depth predictions in foreground human regions, making frame-wise translation refinement less stable than in static-camera scenes.

Multi-HMR 2 (Fiche et al., 2026) is particularly attractive for fast multi-person reconstruction; in multi-person scenes, combining it with scene grounding yields more accurate human–scene reconstruction. For single-person scenes, 3DB (Yang et al., 2026b) already provides an accurate camera-frame translation estimate that can be more reliable than the depth predicted by the scene foundation model. We therefore preserve the original 3DB (Yang et al., 2026b) translation prediction and use its metric human reconstruction as the anchor for scale calibration (Table 4), improving scene scale and camera translation without perturbing the stronger human depth estimate. Thus, scene geometry can refine human motion when the human translation is less reliable, while a strong human model can instead provide a reliable metric anchor for the scene.

Table 10: Ablation of the robust scale estimator for human–scene geometric alignment and global human motion. Both settings use Multi-HMR 2 and VGGT-Ω. Ours denotes our default RANSAC-based calibration, while w/o RANSAC replaces the robust estimator with least-squares fitting. Lower is better for all metrics.

Setting	Human–Scene Alignment				Global Human Motion								
	3DPW		EMDB-1		BEHAVE			RICH			EMDB-2		
	HS-CF5 ↓	HS-CF10 ↓	HS-CF5 ↓	HS-CF10 ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	G-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓
w/o RANSAC	5.614	6.798	3.194	3.428	246.5	188.3	117.3	730.3	289.9	156.3	4865.0	259.7	109.3
Ours	**0.546**	**0.853**	**0.470**	**0.691**	**120.1**	**99.9**	**66.7**	**226.9**	**167.0**	**96.3**	**1580.1**	**210.9**	**78.9**

Across different scene backbones, VGGT-Ω (Wang et al., 2026a) generally provides the most reliable scene geometry and camera poses. Although its native output is scale-normalized and cannot be directly used in a metric world, our calibration maps it into a **near-metric space without retraining**, allowing its stronger geometry to directly benefit world-space human motion reconstruction. In practice, this often makes VGGT-Ω a strong scene backbone for our framework, while Multi-HMR 2 (Fiche et al., 2026) provides efficient multi-person reconstruction and 3DB (Yang et al., 2026b) offers higher-accuracy single-person reconstruction. Overall, our simple geometric alignment improves metric scene and camera reconstruction, enables reliable scene geometry to refine less accurate human motion while preserving already strong human predictions, and allows strong human and scene foundation models to be combined directly without additional joint training.

B.6 Ablation of Robust Metric Calibration

We ablate the robust estimator used in our metric calibration by replacing the default RANSAC-based consensus with least-squares fitting, while fixing Multi-HMR 2 (Fiche et al., 2026) and VGGT-Ω (Wang et al., 2026a) as the human and scene backbones.

As shown in Table 10, least-squares fitting substantially degrades both human–scene geometric alignment and global human motion. Under occlusion or moving-camera settings, where the reconstructed scene can be less reliable, noisy correspondences may be matched to background regions or occluders, leading to erroneous scale estimates or inaccurate human translations. As shown in Figure 11, when the human overlaps with an object, least-squares fitting can be biased toward the nearby object geometry, whereas our RANSAC-based estimator rejects these mismatches and recovers the alignment supported by the dominant human–scene inliers. This highlights the importance of robust consensus estimation for handling noisy human–scene correspondences.

B.7 Qualitative Ablation of Scale Calibration and Translation Refinement

Figures 12 and 13 qualitatively illustrate the effects of scale calibration and translation refinement on BEHAVE (Bhatnagar et al., 2022) (which provides per-frame Kinect-scanned ground-truth scene geometry) as well as internet images.

Effect of Scale Calibration. Figure 12 visualizes metric scene calibration. The left column shows the raw normalized output of VGGT-Ω, whose scale deviates substantially from the ground truth shown in green. After scale calibration (middle), the reconstructed scene is calibrated to a near-metric scale. The right column overlays the raw prediction, our calibrated result, and the ground truth in the same coordinate system, clearly showing the scale correction introduced by our method.

Effect of Translation Refinement. Figure 13 visualizes scene-grounded human translation refinement using Multi-HMR 2 (Fiche et al., 2026). The raw human prediction (left) exhibits a visible gap from the reconstructed scene, whereas translation refinement (middle) updates the human translation to better align the mesh with the scene geometry. The right column directly compares the predictions before and after refinement, with the lighter mesh denoting the raw result. In particular, in the in-the-wild image shown in the last row, the raw predictions fail to establish proper contact between the two people, whereas our refinement brings them into plausible contact while remaining aligned with the reconstructed scene. Overall, translation refinement more accurately grounds the human meshes in the scene and improves geometric consistency in both single- and multi-person interactions.

Table 11: Runtime comparison on a single NVIDIA RTX 4090. End-to-end inference FPS includes image preprocessing and is evaluated on BEHAVE videos at the original resolution of 2048×1536. Core method inference FPS excludes image preprocessing but includes model decoding and subsequent post-processing. For our streaming pipeline, we report both initial and steady-state FPS. Human3R and UniSH do not distinguish between these two stages, and thus a single FPS is reported for each configuration. Higher is better.

Method	Variant	Core Input Resolution		End-to-End Inference FPS ↑		Core Method Inference FPS ↑	
		Human	Scene	Initial	Steady	Initial	Steady
Human3R	672S	672×672	512×384	9.8		13.4	
	672B	672×672	512×384	8.7		11.5	
	896L	896×896	512×384	4.3		4.9	
UniSH	–	192×256	518×392	5.8		6.2	
Ours	Multi-HMR 2+DA3	768×576	504×378	6.2	5.6	8.0	7.2
	Multi-HMR 2+Pi3X	768×576	574×434	5.3	4.6	7.8	6.8
	Multi-HMR 2+VGGT-Ω	768×576	512×384	6.1	5.9	10.0	10.3
	Fast SAM 3D Body+DA3	512×512	504×378	4.4	5.3	4.6	5.8
	Fast SAM 3D Body+Pi3X	512×512	574×434	3.8	4.6	4.5	5.7
	Fast SAM 3D Body+VGGT-Ω	512×512	512×384	4.6	5.7	5.6	7.8

B.8 Runtime Evaluation

We benchmark our method on a single NVIDIA RTX 4090 and re-evaluate Human3R (Chen et al., 2026) and UniSH (Li et al., 2026) on the same compute node used for our experiments, ensuring consistent hardware conditions.

As shown in Table 11, we report both end-to-end and core method inference FPS. End-to-end FPS includes image preprocessing, model inference, decoding, and subsequent post-processing, and is measured on BEHAVE (Bhatnagar et al., 2022) videos at their original resolution of 2048×1536; it is therefore affected by input resolution. Core method FPS excludes only image preprocessing while retaining decoding and post-processing. For our streaming pipeline, we additionally report initial and steady-state FPS, where the initial stage includes the one-time shared-scale initialization, while the steady-state FPS measures subsequent streaming inference after initialization.

Multi-HMR 2 (Fiche et al., 2026) provides the faster configuration, achieving 6.8–10.3 FPS in steady-state core inference, with a human inference cost of about 42 ms/frame. For the higher-accuracy configuration, we use Fast SAM 3D Body (Yang et al., 2026a), an accelerated variant of 3DB (Yang et al., 2026b) that incurs only a small accuracy loss while substantially improving efficiency. All 3DB-based results in this paper use Fast SAM 3D Body, achieving 5.7–7.8 FPS with about 60 ms/frame for human inference.

Runtime also depends on the scene backbone. VGGT-Ω (Wang et al., 2026a) is consistently the most efficient, followed by DA3 (Lin et al., 2026), while Pi3X (Wang et al., 2026b) is the most computationally expensive. By comparison, our geometric alignment adds only about 10 ms/frame during initialization and 4 ms/frame in steady-state inference, showing that the overall runtime is dominated by the pretrained human and scene backbones. Future advances in more efficient human and scene foundation models can therefore directly improve the overall throughput of our pipeline.

B.9 Out-of-Distribution Qualitative Results

We further evaluate COHERE4D on challenging out-of-distribution scenes with severe occlusion and multi-person interaction, as shown in Figures 14 and 15. These examples are drawn from Internet videos together with PROX (Hassan et al., 2019) and Hi4D (Yin et al., 2023).

Robustness to Occlusion. Figure 14 compares human–scene reconstruction under severe occlusion. In the first row, COHERE4D achieves better human–scene alignment and more plausible contact, whereas Human3R places the human above the ground and UniSH shows noticeable mesh–scene misalignment. In the second row, where the lower body is heavily occluded by a table, our method recovers a more plausible occluded-leg configuration while remaining well aligned with the scene; the baselines fail to reconstruct the lower body reliably, and UniSH additionally exhibits

human–scene misalignment. Importantly, this case also highlights the benefit of our training-free design: while unified baselines typically rely on joint training with paired human–scene data, our framework can directly incorporate stronger human and scene foundation models without retraining, so improvements from newer backbones can be immediately combined with our geometric alignment. These results demonstrate both the robustness and flexibility of COHERE4D under strong occlusion.

Multi-Person Interaction. Figure 15 shows challenging multi-person interactions. By grounding the human meshes with foreground scene geometry, COHERE4D better preserves relative human positions and produces more plausible contact between interacting people. Scene grounding also provides additional depth cues for resolving front–back ordering, improving the geometric consistency of multi-person interactions.

B.10 Qualitative Results on In-the-Wild Internet Videos

We further evaluate COHERE4D on diverse Internet videos in Figure 16, including multi-person scenes, dancing, long sequences, and sports motions. Despite the substantial domain shift, COHERE4D maintains accurate human–scene alignment and consistent human–scene geometry over long sequences. These results demonstrate that COHERE4D generalizes robustly beyond benchmark datasets and remains effective for diverse, unconstrained real-world videos.

B.11 Limitations and Failure Cases

Figure 17 illustrates representative failure cases. As discussed in Sec. B.5, our method remains dependent on scene reconstruction quality. Large viewpoint changes or distant humans can degrade the predicted foreground geometry around human regions, reducing the reliability of scene-scale calibration and human translation refinement. As shown on the left of Figure 17, this can occasionally lead to local body–scene misalignment or sparse outliers. Although our pipeline employs a fallback strategy that retains the raw camera-space human translation when the number of reliable inliers is insufficient, a small number of mismatched correspondences may still remain in challenging cases. Among the evaluated scene backbones, VGGT-Ω (Wang et al., 2026a) provides more reliable scene geometry, enabling stronger human translation refinement. As scene foundation models continue to improve, increasingly accurate geometry is expected to further enhance human translation and global motion reconstruction.

Our framework also does not explicitly optimize the tracking capability of the underlying human foundation model. As shown on the right, when multiple people are extremely close, occasional identity association errors can cause overlapping human reconstructions. This limitation is orthogonal to our alignment framework and could be further improved with stronger dense human detection, tracking, or human mesh recovery models.

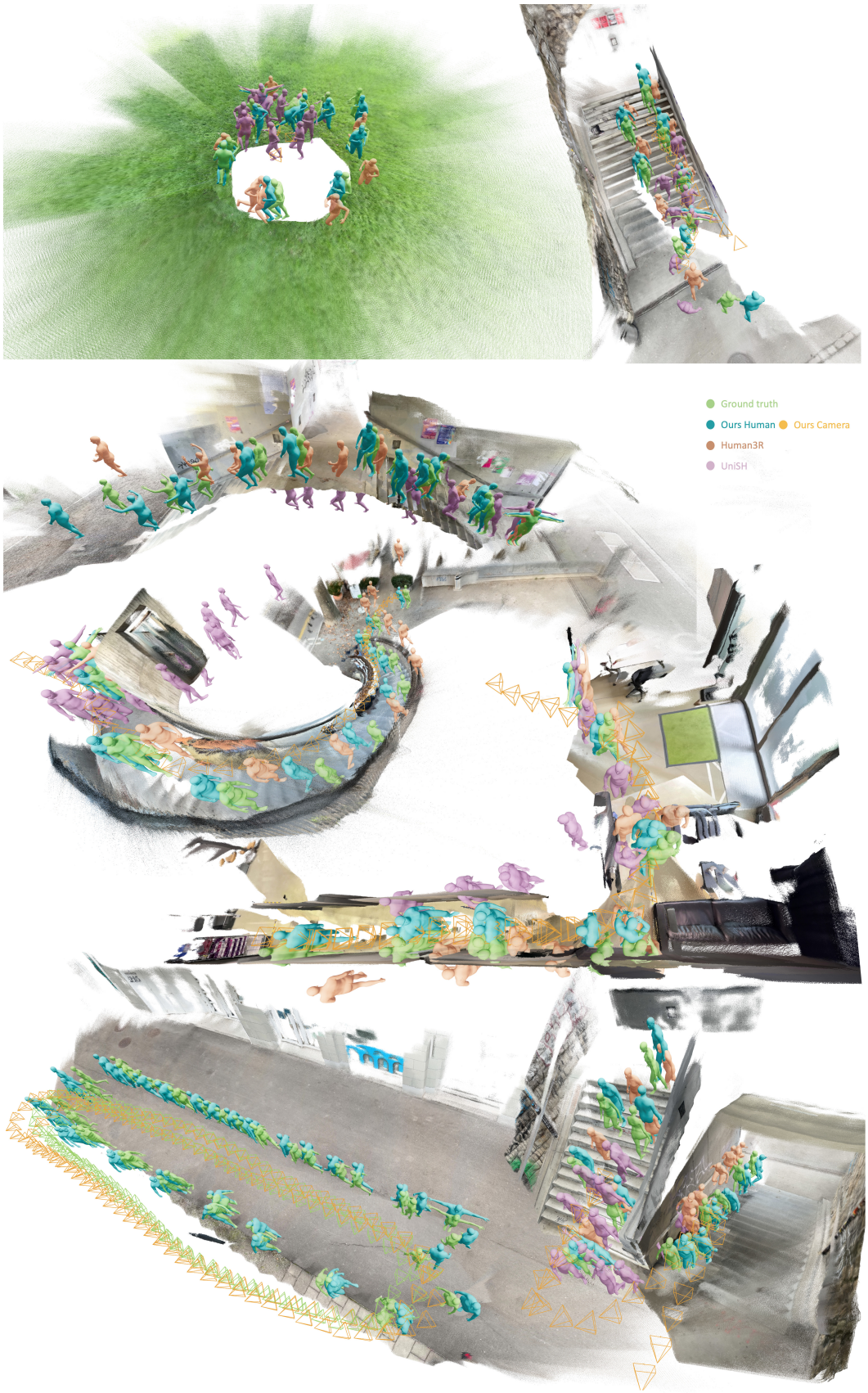

Figure 9: **Qualitative comparison of global human motion** on EMDB-2. COHERE4D exhibits less long-term global trajectory drift and better matches the ground truth than the baselines.

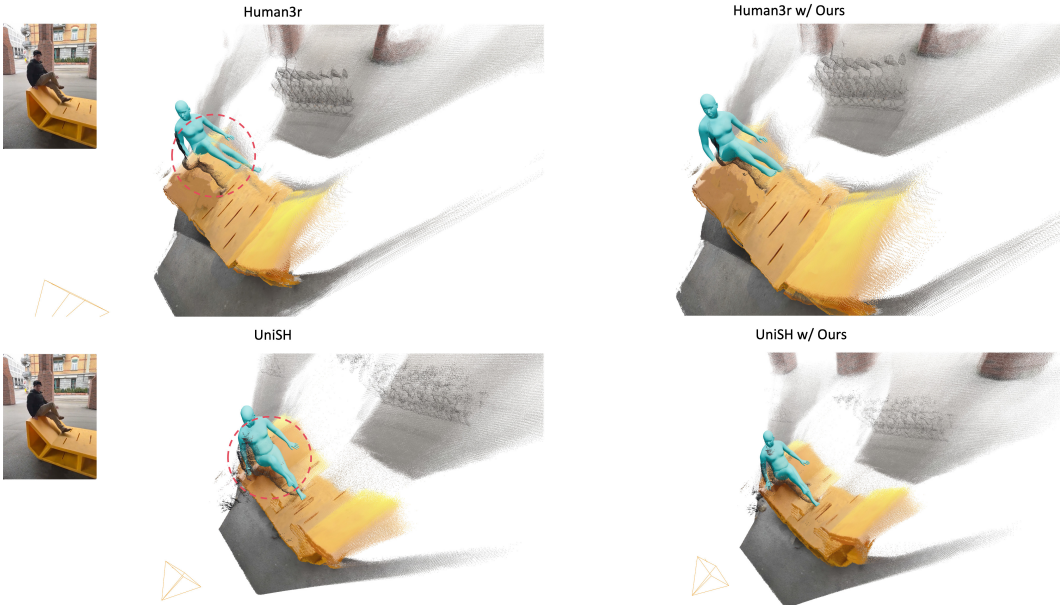

Figure 10: **Improving human–scene alignment in Human3R and UniSH.** Applying CO-HERE4D to either baseline improves human–scene alignment, bringing the human mesh into closer agreement with the reconstructed scene.

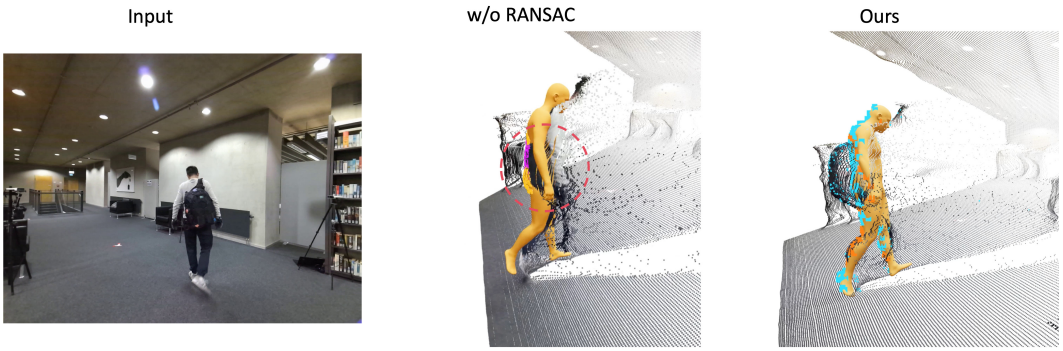

Figure 11: **Qualitative ablation of robust human–scene alignment on BEHAVE.** Without RANSAC, least-squares fitting can be biased by overlapping object geometry, while our robust estimator rejects such mismatches and recovers more accurate human–scene alignment.

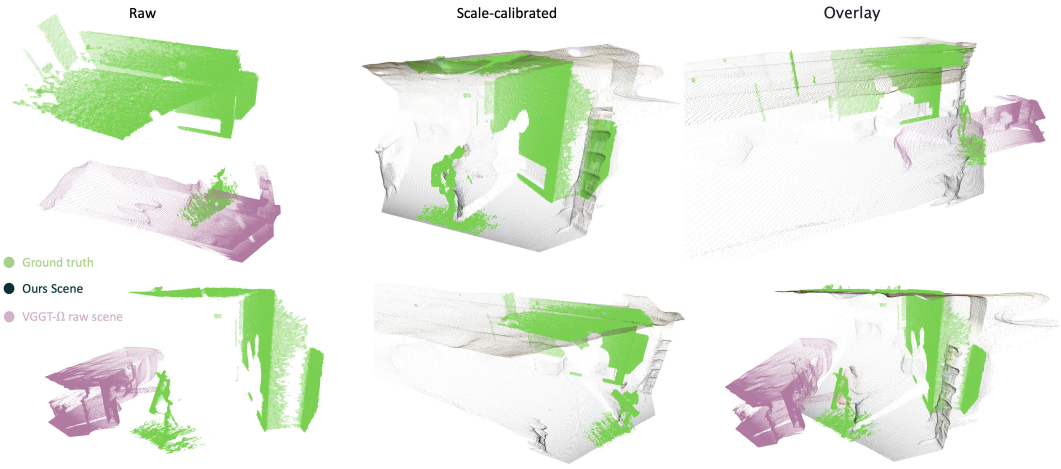

Figure 12: **Qualitative effect of human-anchored scale calibration on BEHAVE.** Our method calibrates the normalized scene prediction toward the ground-truth metric scale.

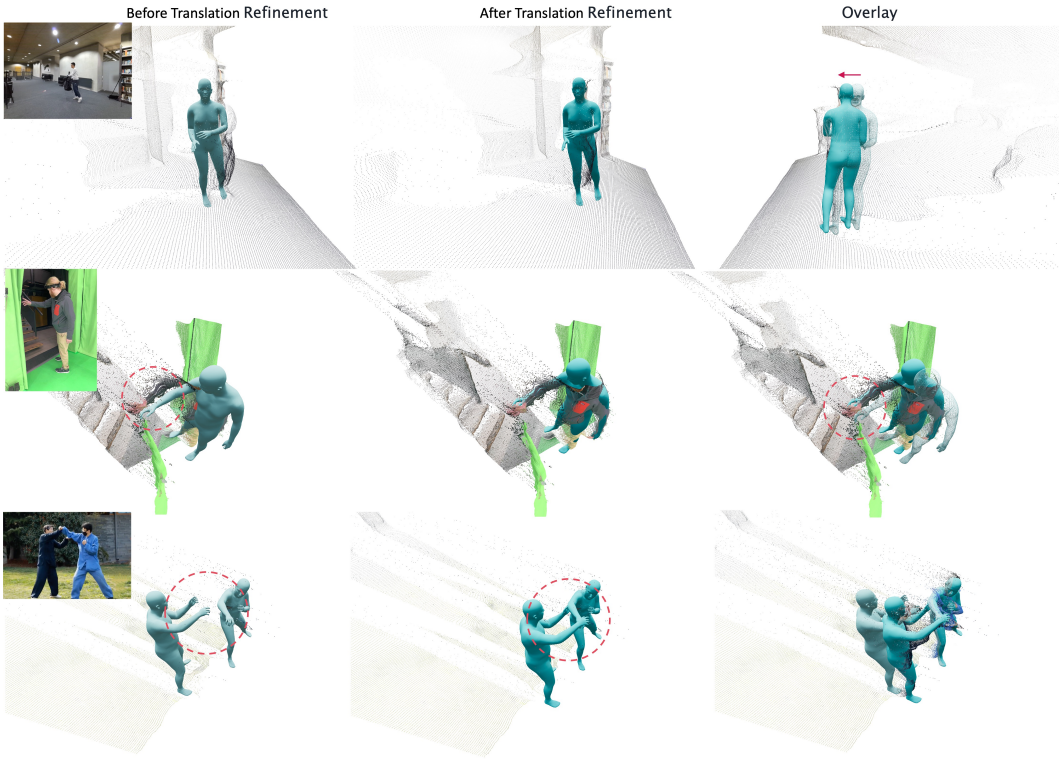

Figure 13: **Qualitative effect of scene-grounded translation refinement.** Our method refines the human translation to better align the mesh with the reconstructed scene. The last row shows that translation refinement improves multi-person contact and interaction.



Figure 14: **Qualitative results under severe occlusion** on Internet videos and PROX. COHERE4D maintains better human–scene alignment and more plausible reconstruction of heavily occluded body parts.



Figure 15: **Qualitative results for multi-person interaction** on Internet videos and Hi4D. Scene grounding improves relative human positioning and produces more plausible contact during interaction.



Figure 16: **Qualitative results on in-the-wild Internet videos.** COHERE4D maintains coherent human–scene alignment and long-term consistency across diverse multi-person, dancing, sports, and long-sequence videos.

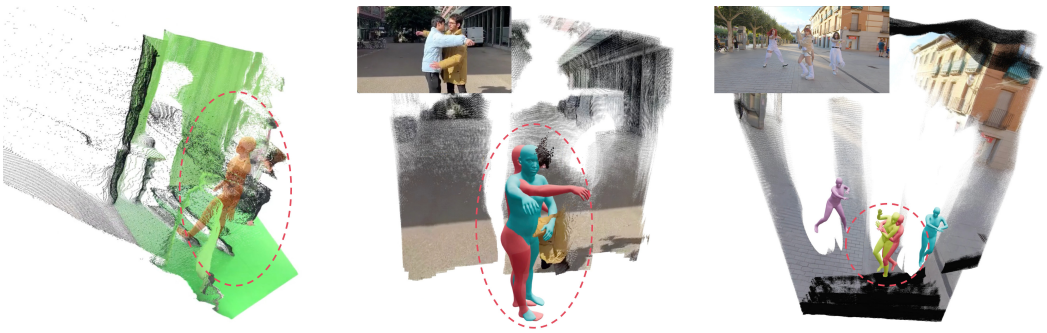

Figure 17: **Representative failure cases of COHERE4D.** Left: unreliable foreground geometry under challenging camera motion can cause occasional alignment errors. Right: close human interactions may lead to identity association errors in the underlying human reconstruction model.