GroundedHuman: Implicit Interaction Grounding for Dynamic Human–Scene Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

Capturing human motion and surroundings from monocular video requires both temporally coherent reconstruction and plausible human-scene alignment. Existing methods typically trade off the two goals: feed-forward models enable efficient dynamic reconstruction but often yield weaker human-scene consistency, whereas interaction-aware methods often target static settings or incur extra overhead. This work presents **GroundedHuman**, a feed-forward framework that reconstructs human motion, scene geometry, and camera motion in a shared metric world. Our key insight is to introduce implicit body-scene interaction grounding into feed-forward reconstruction: the body determines where to inspect the scene geometry, producing interaction representations that enable both metric calibration and motion positioning. Specifically, scene geometry is first calibrated with the metric human body via body-aligned scene attention to establish a shared reference for scene depth and camera motion. Then, interactions between body parts and their surrounding scene context are modeled hierarchically for better positioning of human motions. Experiments show that GroundedHuman improves world-space motion estimation and human-scene coherence while reducing floating and penetration artifacts relative to recent state-of-the-art methods on real-world scenarios.

1 Introduction

Reconstructing humans and their surroundings from monocular video is an important step toward understanding human motion in the 3D world. This requires recovering human motion, scene geometry, and camera motion in a shared metric world while ensuring temporal coherence and plausible human-scene alignment. However, achieving these goals efficiently remains challenging, as reconstructed humans may still float above or penetrate the scene.

Feed-forward methods enable efficient dynamic reconstruction by recovering the human and the scene jointly (Chen et al., 2025; Li et al., 2026), but their human-scene consistency often remains weak. Joint reconstruction methods refine humans, scenes, and cameras through test-time optimization (Liu et al., 2026; Müller et al., 2025). Interaction-aware optimization additionally uses contact and collision constraints to improve human-scene plausibility (Liu et al., 2026; Yalandur Muralidhar et al., 2025). Recent work (Pradyumna et al., 2026) integrates interaction optimization into a feed-forward model for image-based reconstruction, with a visual pipeline that removes humans to recover occluded scene geometry. Meanwhile, there are also attempts (Sur et al., 2026) to explicitly model interaction by inferring contact from human pose and scene geometry, but floating and penetration artifacts remain.

Our key observation is that body geometry and its surrounding spatial context offer complementary information for coherent human–scene reconstruction. Human geometry provides a metric reference for scene dimensions, camera translation, and world-space human motion, while the surrounding spatial context helps determine how the body should be placed within the scene. Based on this observation, we introduce **GroundedHuman**, a feed-forward framework that formulates metric calibration and body positioning as implicit body-scene interaction grounding. The body itself determines where to inspect the scene, and the resulting relations correct both scene scale and body position. We first initialize scene scale from human-scene depth correspondences. We then use body-aligned attention over local scene features to predict residual scale and depth corrections. Aggregating these predictions across people and frames aligns scene geometry with the metric human body, yielding one

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.



Figure 1: (a) GroundedHuman uses metric body queries to inspect scene geometry and jointly recover metric scale and motion position for coherent dynamic human–scene reconstruction. (b) WA-MPJPE$_{100}$ versus inference speed on EMDB-2 (Kaufmann et al., 2023), illustrating the accuracy–efficiency trade-off against recent unified reconstruction methods.

shared calibration of scene depth and camera translation. Within the calibrated scene, we model interactions between body regions and their surrounding scene context hierarchically. We separately encode regional human surface and environmental context across spatial scales and produce predictions of global body position, which are aggregated using reliability weights. After updating body position, we reproject the body and recompute its surrounding context to guide further refinement. This process preserves body pose and shape. Our contributions are summarized as follows:

- We introduce GroundedHuman, a feed-forward framework for monocular dynamic human-scene reconstruction that formulates scene metric calibration and global body positioning as implicit body-scene interaction grounding.
- We propose a body-aligned scene attention mechanism that uses metric body geometry to guide scene metric calibration, establishing a shared metric reference across frames.
- We design a hierarchical context query strategy that models regional human-scene relations as implicit grounding cues for interaction-aware global body positioning.

2 Related Work

3D/4D Scene Reconstruction Recent visual geometry foundation models have enabled direct prediction of scene geometry and camera motion from RGB observations. DUSt3R (Wang et al., 2024a) estimates pixel-aligned 3D pointmaps, while MASt3R (Leroy et al., 2024) introduces dense feature matching to establish cross-view correspondences. VGGT (Wang et al., 2025a) jointly predicts camera parameters, depth, and dense geometry, whereas π^3 (Wang et al., 2026b) achieves reference-free reconstruction through permutation-equivariant geometric prediction. For dynamic and sequential inputs, MonST3R (Zhang et al., 2025) models geometry in the presence of motion, CUT3R (Wang et al., 2025b) maintains a persistent state for continuous scene reconstruction, and VGGT-Ω (Wang et al., 2026a) advances feed-forward geometry estimation for both static and dynamic scenes. D4RT (Zhang et al., 2026) unifies dynamic reconstruction and tracking through point queries over a global scene representation, motivating our body-conditioned scene querying. Despite these advances, monocular scale ambiguity and motion-induced occlusion and depth discontinuities can hinder metric alignment between scene geometry and independently reconstructed humans.

Human Reconstruction and Global Motion Human Mesh Recovery (HMR) estimates 3D human pose and shape, commonly using parametric body models such as SMPL (Loper et al., 2023) and SMPL-X (Pavlakos et al., 2019), with methods such as HMR 2.0 (Goel et al., 2023), Multi-HMR (Baradel et al., 2024), and NLF (Sárándi & Pons-Moll, 2024) recovering expressive or dense human representations from monocular images. Beyond local body reconstruction, world-grounded HMR recovers human motion in a shared world coordinate frame. GLAMR (Yuan et al., 2022) uses learned motion priors to recover global trajectories under occlusion, while SLAHMR (Ye et al., 2023) combines relative camera motion with human motion priors for global reconstruction. To resolve metric ambiguity, TRAM (Wang et al., 2024b) uses SLAM and scene-background cues to recover global scale, while WHAC (Yin et al., 2024) exploits absolute human depth and motion cues

to jointly recover metric human and camera trajectories. WHAM (Shin et al., 2024) combines motion context with contact-aware trajectory refinement, and GVHMR (Shen et al., 2024) introduces gravity-view coordinates for world-grounded motion recovery. More recently, DuoMo (Wang et al., 2026c) separates camera-space and world-space motion generation into two diffusion stages. Nevertheless, temporally coherent motion does not guarantee accurate human position without sufficient constraints from scene geometry.

Scene-Human Reconstruction Recent methods combine human priors and scene geometry for joint reconstruction. SynCHMR (Zhao et al., 2024) uses human-aware metric SLAM to recover camera poses and scene geometry, followed by scene-aware motion refinement. MetricHMSR (Song et al., 2026) incorporates camera-ray representations for metric human recovery and uses reconstructed human surfaces as geometric anchors to refine monocular scene depth. JOSH (Liu et al., 2026) jointly optimizes camera poses, scene geometry, and human motion through correspondence and contact constraints. Beyond geometric priors and joint optimization, recent feed-forward methods integrate human and scene reasoning directly into reconstruction networks. HAMSt3R (Rojas et al., 2025) augments multi-view reconstruction with human segmentation and dense body correspondences, while Human3R (Chen et al., 2025) incorporates human prompts into a recurrent geometry model to jointly recover scene structure, camera motion, and global human meshes. UniSH (Li et al., 2026) combines scene reconstruction and HMR priors with explicit human-scene geometric correspondence for joint metric-scale reconstruction. SHOW (Shi et al., 2026) introduces bidirectional geometric conditioning: human semantics and body-scale priors guide scene reconstruction, while scene features constrain human mesh recovery and global localization. To further address floating and interpenetration artifacts, UniCon3R (Sur et al., 2026) feeds predicted human-scene contact back into human reconstruction as a corrective signal, while GRAFT (Pradyumna et al., 2026) employs body-anchored geometric probes and recurrent refinement to adjust human meshes according to nearby scene surfaces. These methods highlight the value of geometric feedback, motivating global position refinement that preserves reliable local body pose and shape.

3 Method

Given a monocular video, our goal is to reconstruct scene geometry, camera motion, and human meshes in a shared metric world coordinate system. The overview of our method is illustrated in Fig. 2.

3.1 Problem Formulation

We start from frozen VGGT-Ω (Wang et al., 2026a) and NLF (Sárándi & Pons-Moll, 2024) as the scene and human estimators, respectively, which provide complementary geometric observations. The scene estimator predicts relative scene depth, camera parameters, and multi-layer visual features, while the human estimator recovers metric body pose, shape, root translation, and confidence for each person. These observations provide the human structure and scene evidence used by the following grounding modules.

We formulate geometric correction as implicit grounding between human and scene representations. Human structure determines where and how scene evidence is retrieved, while the resulting human-scene relationship is encoded directly to predict the required correction. Formally,

$$\mathcal{Y} = \mathcal{G}(\mathcal{H}, \mathcal{S}), \tag{1}$$

where $\mathcal{H}$ denotes the human representation, $\mathcal{S}$ denotes its associated scene evidence, and $\mathcal{G}$ denotes the implicit grounding function that predicts a geometric correction $\mathcal{Y}$ from their relationship.

GroundedHuman instantiates this formulation through two complementary grounding modules. **Body-Aligned Metric Calibration** (Sec. 3.2) grounds relative scene geometry to metric human observations, recovering a shared metric reference for scene depth and camera motion. Given the calibrated scene, **Hierarchical Interaction Grounding** (Sec. 3.3) grounds human body regions against hierarchical scene context to improve interaction-consistent positioning.

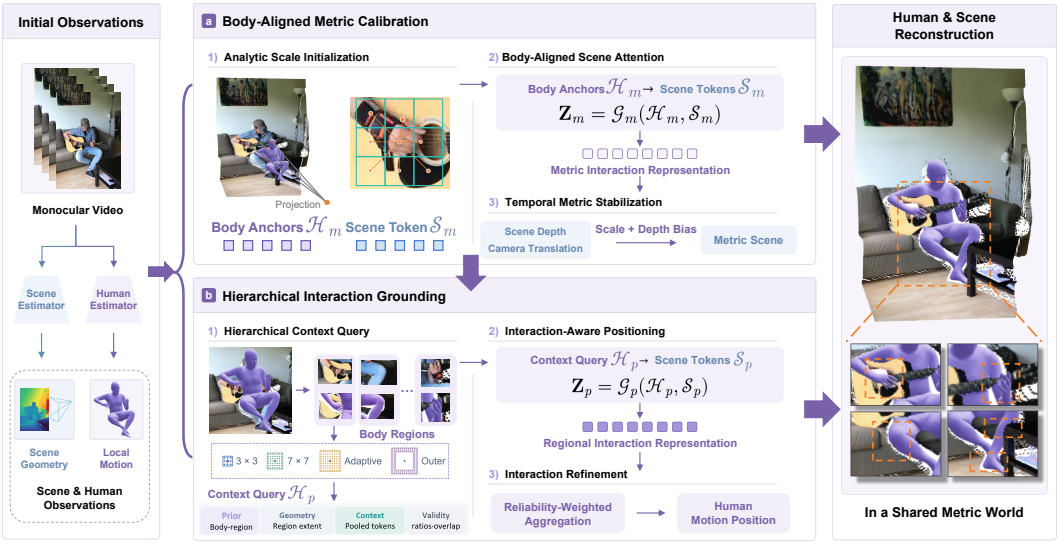


Figure 2: Overview of GroundedHuman. (a) Body-Aligned Metric Calibration uses human anchors to query local scene evidence and calibrate scene depth and camera translation. (b) Hierarchical Interaction Grounding further queries scene geometry around body regions to refine human motion position within the calibrated scene.

3.2 Body-Aligned Metric Calibration

We instantiate $\mathcal{G}$ in Eq. 1 for metric calibration as a body-aligned grounding module, which queries local scene evidence around metric human anchors and predicts residual corrections to scene scale and depth. The module consists of three components: Analytic Scale Initialization, Body-Aligned Scene Attention, and Temporal Metric Stabilization.

Analytic Scale Initialization. Let D_t^r and $\mathbf{K}_t$ denote the relative scene depth and camera intrinsics at frame t, and $\mathbf{V}_t$ is the initial body vertices in camera coordinates. To establish a coarse metric grounding between the body and scene, we associate visible body vertices with scene depths along the same camera rays and estimate their proportional scale:

$$s_t^c = \Phi_{\text{scale}}(\mathbf{V}_t, \mathbf{K}_t, D_t^r)\,. \tag{2}$$

Here, Φ_{scale} establishes valid body-scene depth correspondences and aggregates them into a frame-level scale. The coarse depth is then initialized as $D_t^{\text{coarse}} = s_t^c D_t^r$. Implementation details including projection, visibility filtering, aggregation, and missing-observation are provided in Appendix B.

Body-Aligned Scene Attention. Depth ratios alone cannot distinguish scale mismatch from placement or scene-depth errors. We use N_a body anchors per person to interpret residual discrepancies using body geometry and local scene context. Each anchor token encodes the body state, anchor position and projection, together with nearby geometric observations reconstructed from D_t^{coarse}. The resulting collection of body-anchor tokens is denoted by $\mathbf{H}_t$. Around each projected anchor, a 3×3 neighborhood is sampled from the multi-layer scene features $\mathbf{F}_t$ predicted by the scene estimator, forming local scene tokens $\mathbf{S}_t$.

We first apply self-attention to $\mathbf{H}_t$ to capture body-wide context across anchors, producing contextualized body tokens $\widetilde{\mathbf{H}}_t$. These tokens then serve as queries in cross-attention over the local scene tokens $\mathbf{S}_t$, yielding scene-conditioned features $\mathbf{Z}_t$ that retrieve nearby scene evidence. The attended features are aggregated into a body-level representation $\mathbf{z}_t = \text{Pool}(\mathbf{Z}_t)$. Two lightweight prediction heads then estimate the residual scale correction and depth bias:

$$\Delta s_t = \text{MLP}_s(\mathbf{z}_t), \qquad \Delta b_t = \text{MLP}_b(\mathbf{z}_t). \tag{3}$$

Temporal Metric Stabilization. Individual predictions remain sensitive to visibility and estimation errors. We therefore aggregate the coarse scales and predicted corrections across people and

time, using confidence scores c_t provided by the human estimator:

$$(s, b) = \Phi_{\text{temp}}\left(\{s_t^c, \Delta s_t, \Delta b_t, c_t\}_{t=1}^{T}\right), \tag{4}$$

where Φ_{temp} denotes robust confidence-aware temporal metric stabilization. Its implementation is detailed in Appendix B. To keep scene geometry and camera motion in the same metric reference, the resulting scale is applied consistently to both scene depth and relative camera translation $\mathbf{t}_t^r$:

$$D_t^m = sD_t^r + b, \qquad \mathbf{t}_t^m = s\mathbf{t}_t^r. \tag{5}$$

The additive bias is applied only to depth, while camera rotation and intrinsics remain unchanged. Back-projecting D_t^m with $\mathbf{K}_t$ yields camera-space metric points $\mathbf{P}_t^m$, which provide the reference for human motion position refinement.

3.3 Hierarchical Interaction Grounding

Metric calibration establishes a common metric reference, but inaccurate human position can still cause floating or interpenetration. We therefore refine humam motion position by querying scene geometry around body regions. The key idea is to combine local surface alignment with broader environmental context and update this context as the body moves. Frame index t is omitted for clarity.

Hierarchical Context Query. We partition the SMPL surface into N_r corresponding regions, each represented by its centroid and eight sampled vertices. To capture both surface alignment and environmental support, each region gathers scene evidence from four supports, denoted collectively by $\mathcal{W}$: fixed 3×3 and 7×7 windows, an adaptive window following the projected region extent, and an outer annulus.

For each support, we construct two query tokens encoding the body-region state and geometric context, yielding eight region queries $\mathbf{Q}_r$. Scene tokens $\mathbf{S}_r$ are sampled within the corresponding supports from the multi-layer features $\mathbf{F}$ predicted by the scene estimator. Cross-attention uses $\mathbf{Q}_r$ as queries and $\mathbf{S}_r$ as keys and values under a support-aware mask, producing $\mathbf{Z}_r$, which are fused with regional cues into the interaction representation $\mathbf{z}_r$. Details are provided in Appendix B.

Interaction-Aware Positioning. For each region, we define a local translation basis consisting of the viewing direction $\mathbf{e}_{\parallel,r}$ and two orthogonal directions $\mathbf{e}_{1,r}$ and $\mathbf{e}_{2,r}$ spanning the perpendicular plane. A lightweight prediction head maps the interaction representation $\mathbf{z}_r$ to the corresponding displacement coefficients $\boldsymbol{\delta}_r = (\delta_{\parallel,r}, \delta_{1,r}, \delta_{2,r})^\top$, together with a gate g_r and scalar log-variance ℓ_r:

$$(\boldsymbol{\delta}_r, g_r, \ell_r) = \text{MLP}_{\text{position}}(\mathbf{z}_r). \tag{6}$$

The regional translation proposal is then expressed as

$$\mathbf{v}_r = \delta_{\parallel,r}\mathbf{e}_{\parallel,r} + \delta_{1,r}\mathbf{e}_{1,r} + \delta_{2,r}\mathbf{e}_{2,r}. \tag{7}$$

Because the available evidence varies in reliability across regions, we robustly aggregate the regional proposals using their predicted gates and uncertainties:

$$\widehat{\mathbf{v}} = \Phi_{\text{agg}}\left(\{\mathbf{v}_r, g_r, \ell_r\}_{r=1}^{N_r}\right), \tag{8}$$

where Φ_{agg} suppresses invalid regions and down-weights uncertain predictions. The detailed weighting scheme is provided in Appendix B. Translation supervision directly trains the regional representation and prediction head without explicit contact labels.

Interaction Refinement. Let $\boldsymbol{\tau}$ denote the current root translation. To limit unreliable corrections, we clip the aggregated proposal to obtain the interaction-aware root update $\Delta\boldsymbol{\tau}$. The translation is updated as

$$\boldsymbol{\tau} \leftarrow \boldsymbol{\tau} + \Delta\boldsymbol{\tau}. \tag{9}$$

We perform two feed-forward refinement iterations with shared parameters. Since each translation update changes the human-scene interaction configuration, regional queries and their associated scene context are recomputed before the next correction. Scene geometry, camera parameters, body pose, and shape remain fixed throughout refinement. Finally, calibrated inverse camera extrinsics map the refined humans and metric scene points into the shared world coordinate system (Appendix B). Implementation details, training strategies, datasets, and loss functions for both modules are provided in Appendix B.3.

4 EXPERIMENTS

We evaluate global human motion on EMDB-2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022), scene depth on Bonn (Palazzolo et al., 2019), camera trajectories on the dynamic sequences of the TUM RGB-D benchmark(TUM-Dynamic) (Sturm et al., 2012), and human-scene consistency on 3DPW (Von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023); local human mesh results are provided in Appendix A.2. Qualitative comparisons on PROX-D (Hassan et al., 2019) and outdoor sequences examine human-scene alignment, while measurements on PROX (Hassan et al., 2019) assess human and environmental dimensions. Implementation details are provided in Appendix B.3.

Table 1: Evaluation of global human motion estimation on EMDB-2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022). Contact Control denotes a human-scene grounding mechanism that adjusts human placement to reduce interpenetration, through explicit constraints or learned geometric updates; it does not require explicit contact-label prediction.

Category	Method	Capabilities			EMDB-2 (Kaufmann et al., 2023) (24)			RICH (Huang et al., 2022) (24)		
		Multi-Person	Metric human-scene	Interaction-Aware	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓
Offline	GLAMR (Yuan et al., 2022)	✓	×	×	280.8	726.6	11.4	129.4	236.2	3.8
	SLAHMR (Ye et al., 2023)	✓	×	×	326.9	776.1	10.2	132.2	237.1	6.4
	COIN (Li et al., 2024)	×	×	×	152.8	407.3	3.5	169.5	254.5	–
	GVHMR (Shen et al., 2024)	×	×	×	111.0	276.5	2.0	78.8	126.3	2.4
	TRAM (Wang et al., 2024b)	×	×	×	76.4	222.4	1.4	127.8	238.0	6.0
	JOSH (Liu et al., 2026)	✓	✓	✓	68.9	174.7	1.3	89.0	132.5	3.0
	PromptHMR (Wang et al., 2025c)	✓	×	×	71.0	216.5	1.3	118.2	207.0	5.6
Online	TRACE (Sun et al., 2023)	✓	×	×	529.0	1702.3	17.7	238.1	925.4	610.4
	WHAM (Shin et al., 2024)	×	×	×	135.6	354.8	6.0	108.4	196.1	4.5
	JOSH3R (Liu et al., 2026)	×	✓	×	220.0	661.7	13.1	–	–	–
	Human3R (Chen et al., 2025)	✓	✓	×	112.2	267.9	2.2	110.0	184.9	3.3
	MetricHMSR (Song et al., 2026)	×	✓	×	72.1	199.5	1.4	109.6	165.5	1.9
	UniCon3R (Sur et al., 2026)	×	✓	✓	113.7	285.6	2.3	**81.5**	129.8	**1.7**
	UniSH (Li et al., 2026)	×	✓	×	118.5	270.1	5.8	118.1	183.2	4.8
	SHOW (Shi et al., 2026)	×	✓	×	109.1	262.3	2.1	107.3	172.7	2.2
	Ours	✓	✓	✓	**58.5**	**150.0**	**1.1**	87.7	**128.2**	2.5

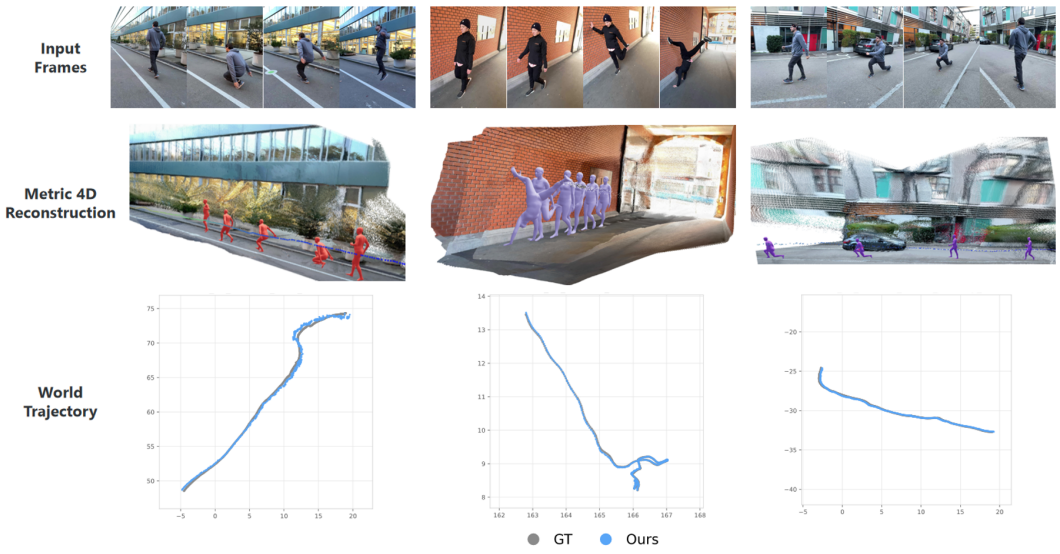
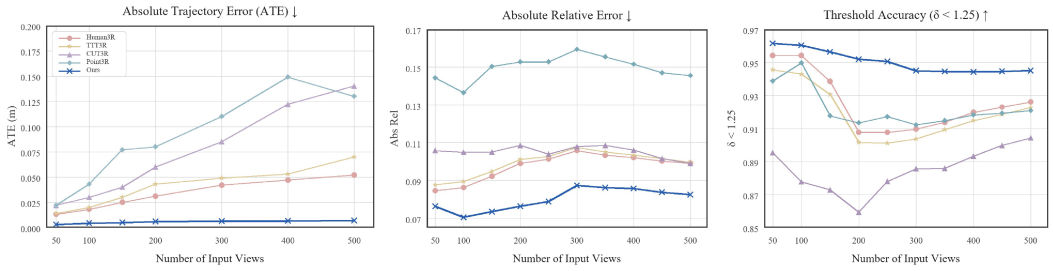


Figure 3: Temporal human-scene reconstruction. Each column shows input frames, reconstructed bodies within the scene, and world trajectories. Gray: ground truth; blue: ours.

4.1 GLOBAL HUMAN MOTION ESTIMATION

We assess global human motion on EMDB-2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022) using WA-MPJPE, W-MPJPE, and RTE. The joint metrics assess world-space motion under different alignment scopes, while RTE measures root-trajectory accuracy.

World-space human motion depends on both camera motion and scene scale. Our comparison includes dedicated motion estimators and joint reconstruction methods (Table 1). GroundedHuman



(a) Camera pose estimation. (b) Video depth estimation in metric scale.

Figure 4: Evaluation of generic 3D reconstruction across different input lengths. (a) Camera pose estimation on TUM-Dynamic (Sturm et al., 2012) using absolute trajectory error (ATE) after Sim(3) alignment, evaluating trajectory accuracy up to a global similarity transform rather than metric-scale accuracy. (b) Video depth estimation on Bonn (Palazzolo et al., 2019) in metric scale using Abs Rel and $\delta < 1.25$. For GroundedHuman, metric depth is obtained with the Body-Anchored Metric Calibration module in Sec. 3.2, without per-sequence ground-truth scale or shift fitting.

places the body sequence within its environment through a scale shared by scene depth and camera translation.

The reconstructed sequences expose articulation and global displacement within the same scene (Fig. 3). Sports warm-up, handstand, and walking-with-lunges examples show bodies at successive instants, alongside predicted and reference world trajectories. This joint view relates each motion path to the surrounding geometry rather than presenting it as an isolated trajectory.

4.2 Scene Reconstruction

Camera Pose Estimation. Figure 4(a) evaluates camera trajectories on TUM-Dynamic (Sturm et al., 2012) using ATE after Sim(3) alignment, as also used in Human3R (Chen et al., 2025). The comparison covers different input lengths and representative online reconstruction models: Human3R (Chen et al., 2025), TTT3R (Chen et al., 2026), CUT3R (Wang et al., 2025b), and Point3R (Wu et al., 2026). Since Sim(3) alignment absorbs the uniform scale applied to camera translation, this evaluation measures trajectory accuracy up to a global similarity transform and does not isolate the contribution of metric calibration. Metric depth accuracy is evaluated separately in Fig. 4(b), without ground-truth scale or shift fitting.

Video Depth Estimation. Figure 4(b) evaluates video depth on Bonn (Palazzolo et al., 2019) using Abs Rel and $\delta < 1.25$, which measure relative depth error and the fraction of predictions within a factor of 1.25 of the ground truth, respectively. The predicted relative depth is converted to metric scale by the Body-Anchored Metric Calibration module in Sec. 3.2: human surfaces initialize scale, Body-Anchored Scene Attention encodes correspondence for a metric readout that predicts residual scale and depth bias, and temporal aggregation applies one shared calibration across the video. No per-sequence ground-truth scale or shift is fitted for these depth results. Across input lengths, the resulting metric depth remains more accurate than the compared online baselines on both measures.

Additional local human mesh results are reported in Appendix A.2, including the PA-MPJPE, MPJPE, and PVE comparison on 3DPW (Von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023). The appendix retains the complete comparison table and its evaluation description before the main-paper discussion of human-scene consistency.

4.3 Human-Scene Consistency

We evaluate human-scene consistency on 3DPW (Von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023) using human-scene Variance (HS-V) and human-scene Chamfer (HS-CF), introduced by SHOW (Shi et al., 2026). HS-V measures the discrepancy between the horizontal and vertical spatial variances of the visible body surface and the scene point cloud within the human mask. HS-CF measures their normalized nearest-neighbor surface distance. Subscripts 5 and 10 denote the source protocol's 5th–95th and 10th–90th percentile filtering ranges, respectively. These

Table 2: human-scene consistency on 3DPW (Von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023). Baseline results are taken from SHOW (Shi et al., 2026). Lower is better for all metrics; bold and underlined values denote the best and second-best reported results, respectively.

Method	3DPW (Von Marcard et al., 2018)				EMDB-1 (Kaufmann et al., 2023)			
	HS-V_5 ↓	HS-V_{10} ↓	HS-CF_5 ↓	HS-CF_{10} ↓	HS-V_5 ↓	HS-V_{10} ↓	HS-CF_5 ↓	HS-CF_{10} ↓
Human3R (Chen et al., 2025)	0.652	0.640	**1.30**	**1.32**	3.66	2.91	**1.13**	**1.31**
UniSH (Li et al., 2026)	<u>0.028</u>	0.028	7.65	7.52	0.033	0.037	7.10	6.99
SHOW (Shi et al., 2026)	0.030	<u>0.026</u>	5.69	5.44	<u>0.013</u>	**0.017**	5.88	5.79
Ours	**0.0155**	**0.0181**	<u>4.99</u>	<u>4.86</u>	**0.0102**	<u>0.0181</u>	<u>3.75</u>	<u>3.67</u>

metrics assess agreement between the reconstructed body and its observed surface in the scene representation, complementing the collision and floating diagnostics used in the following ablation.

Table 2 shows that GroundedHuman improves all four metrics over UniSH (Li et al., 2026) on both datasets. On 3DPW (Von Marcard et al., 2018), it achieves the lowest HS-V_5/HS-V_{10} (0.0155/0.0181), while reducing HS-CF_5/HS-CF_{10} from SHOW's (Shi et al., 2026) 5.69/5.44 to 4.99/4.86. On EMDB-1 (Kaufmann et al., 2023), it achieves the lowest HS-V_5 (0.0102) and reduces HS-CF_5/HS-CF_{10} from 5.88/5.79 to 3.75/3.67; SHOW (Shi et al., 2026) retains a slightly lower HS-V_{10} (0.017 versus 0.0181). Human3R (Chen et al., 2025) achieves the lowest HS-CF on both datasets but has substantially larger HS-V. Together, these results show that GroundedHuman combines low spatial-extent discrepancy with competitive visible-surface alignment. They assess the consistency of the complete reconstruction pipeline; the following ablation isolates the contribution of Implicit Interaction Grounding.

4.4 Ablation of Implicit Interaction Grounding

Table 3 ablates Implicit Interaction Grounding and its four components. All variants retain metric calibration and preserve body pose and shape. We follow the RICH (Huang et al., 2022) physical-grounding protocol of UniCon3R (Sur et al., 2026), including robust ground-height estimation and a 5 mm tolerance, with identical evaluation samples, metric denominators, and aggregation rules. We report collision ratio (Coll.), penetration (Pen.), floating distance (Float), and maximum penetration (P.Max).

GroundedHuman improves all four metrics over disabling grounding, reducing Coll. from 12.83% to 6.48%. Removing context re-encoding or self/environment separation worsens every metric, supporting placement-dependent context updates and separate evidence channels. Removing multi-scale context lowers Pen. to 2.81 cm but worsens the other three metrics; removing reliability weighting lowers Float to 13.26 cm but likewise worsens the remaining metrics. The complete model achieves the lowest Coll. and P.Max and the second-lowest Pen. and Float, balancing penetration and floating errors.

Table 3: Ablation of Implicit Interaction Grounding on RICH (Huang et al., 2022) under the physical-grounding protocol of UniCon3R (Sur et al., 2026). Each ablated variant removes the indicated component from GroundedHuman while retaining metric calibration. Lower is better; bold and underlined values denote the best and second-best results, respectively.

Variant	Coll. (%) ↓	Pen. (cm) ↓	Float (cm) ↓	P.Max (cm) ↓
w/o Implicit Interaction Grounding	12.83	4.65	17.15	7.99
w/o Context Re-encoding	<u>7.46</u>	3.10	14.02	<u>7.39</u>
w/o Multi-scale Context	8.31	**2.81**	15.18	7.58
w/o Self/Environment Separation	9.42	3.71	15.56	7.76
w/o Reliability Weighting	8.88	3.49	**13.26**	8.14
GroundedHuman	**6.48**	<u>2.89</u>	<u>13.47</u>	**7.31**

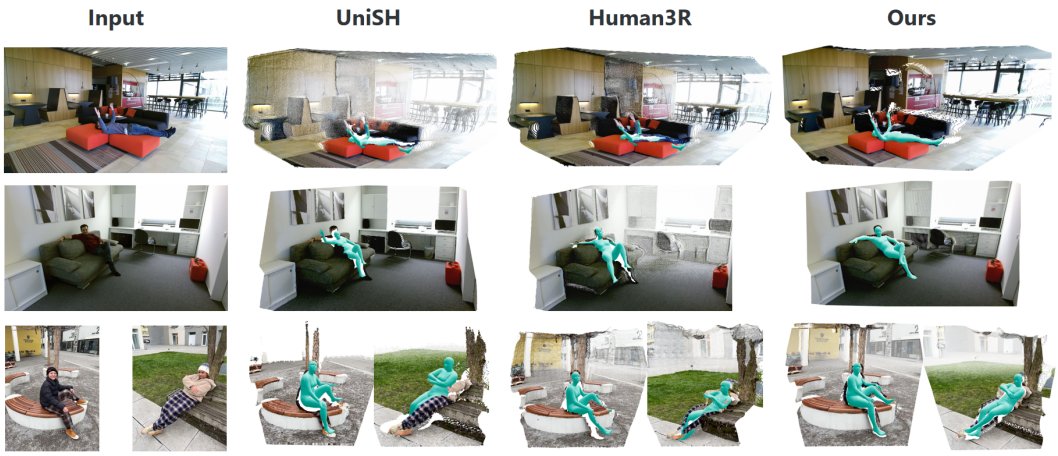


Figure 5: Human placement in reconstructed scenes. Columns show the input, UniSH (Li et al., 2026), Human3R (Chen et al., 2025), and GroundedHuman. Lying, reclining, and outdoor sitting examples highlight the position of the body relative to nearby support surfaces.

4.5 Qualitative Results

Human-Scene Spatial Alignment. We compare human placement with UniSH (Li et al., 2026) and Human3R (Chen et al., 2025) on PROX-D (Hassan et al., 2019) and outdoor sequences containing lying, reclining, and sitting poses (Fig. 5). The indoor scenes emphasize torso-to-cushion and pelvis-to-seat alignment, while the outdoor scenes expose the position of the body relative to benches and backrests. These relationships reveal global offsets that can remain inconspicuous when a body is inspected in isolation. Our regional refinement uses the surrounding geometry to adjust root position while preserving the observer's local articulation.

5 Conclusion

We presented GroundedHuman, a feed-forward framework that reconstructs human motion, scene geometry, and camera motion in a shared metric world. We introduce implicit body-scene interaction grounding into feed-forward reconstruction: the body determines where to inspect the scene geometry, producing interaction representations that enable both metric calibration and motion positioning. Experiments on EMDB, RICH, and 3DPW show that GroundedHuman improves world-space motion estimation and human-scene coherence while reducing floating and penetration artifacts relative to recent state-of-the-art methods.

Limitation and Future Work. GroundedHuman recovers the metric scale of the scene from the human body, which makes the recovered scale dependent on the accuracy of monocular human estimation, where body shape and root position are only approximate. The framework performs feed-forward inference, but it is not yet fast enough for real-time applications. Penetration artifacts also remain in challenging cases. Improving the robustness of the metric estimate, speeding up inference, and reducing these remaining penetration artifacts are directions for future work.

Acknowledgments

Omitted in the double-blind review version.

6 AI Use Statement

LLMs were used solely for grammar checking in the preparation of this manuscript. They were not employed in any research activities or in drafting the scientific content.

7 Ethics Statement

This work adheres to the ICLR Code of Ethics. Our research is based exclusively on publicly accessible datasets and does not include human subjects, personally identifiable information, or sensitive data. The methods proposed herein are designed solely for the advancement of fundamental machine learning research.

8 Reproducibility Statement

To facilitate reproducibility, we provide detailed descriptions of the model architecture, training settings, and evaluation protocols in both the main paper and the supplementary material.

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.

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.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025.

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

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.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13264–13275. IEEE, 2022.

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

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

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

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, et al. Unish: Unifying scene and human reconstruction in a feed-forward pass. *arXiv preprint arXiv:2601.01222*, 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.

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

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.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

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

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

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems*, 37:140032–140065, 2024.

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.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

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.

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

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. 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, 2025c.

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

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *Advances in Neural Information Processing Systems*, 38:69675–69699, 2026.

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

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.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pp. 20–37. Springer, 2024.

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. 11028–11039. IEEE, 2022.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one d4rt at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2026.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, volume 2025, pp. 82863–82886, 2025.

Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-space camera and human reconstruction from videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1216–1226. IEEE, 2024.

A Metric Definitions and Evaluation Protocols

A.1 Global Human Motion

For a 100-frame segment, let $\widehat{\mathbf{J}}_{t,k}$ and $\mathbf{J}_{t,k}$ denote predicted and reference world-space joints. WA-MPJPE fits one similarity transform on all valid joints in the segment and reports the mean Euclidean joint error in millimeters:

$$\text{WA-MPJPE} = \frac{1000}{|\mathcal{V}|} \sum_{(t,k)\in\mathcal{V}} \left\| \mathcal{S}_{\text{seg}}(\widehat{\mathbf{J}}_{t,k}) - \mathbf{J}_{t,k} \right\|_2 .$$

W-MPJPE estimates the similarity transform from the first two valid frames and applies that transform to the entire segment. It therefore exposes long-horizon trajectory drift that a segment-wide alignment can hide. Root Translation Error uses the root trajectories after rigid alignment and normalizes the mean error by the reference root displacement:

$$\text{RTE}(\%) = 100 \, \frac{\sum_t \|\widehat{\mathbf{r}}_t - \mathbf{r}_t\|_2}{\sum_t \|\mathbf{r}_t - \mathbf{r}_{t-1}\|_2 + \epsilon}.$$

Missing detections are excluded from the matched set and the valid-frame coverage is recorded with the evaluation output.

A.2 Local Human Mesh Recovery

For camera-space joints, MPJPE is the mean Euclidean error after pelvis translation removal. PA-MPJPE additionally applies a per-frame Procrustes similarity alignment, so it emphasizes articulated pose and shape agreement. PVE is the mean Euclidean error over corresponding mesh vertices.

A.3 Human–Scene Consistency: Table 2

We use the human–scene consistency metrics introduced by SHOW (Shi et al., 2026). For each frame, the SMPL mesh is projected into the camera image to construct the human mask used to select scene-depth observations. The visible human surface is obtained by projecting the SMPL mesh and resolving visibility in the camera view. Invalid scene-depth samples are removed, and boundary outliers are filtered by either the 5th–95th or the 10th–90th percentile rule. The subscript in HS-CF_5/HS-V_5 or HS-CF_{10}/HS-V_{10} records that filtering choice. All quantities are computed in the camera frame before averaging over valid frames.

Let $H = \{\mathbf{h}_i\}_{i=1}^N$ be visible human points and let S_α be the filtered scene point set. The normalized Human–Scene Chamfer is

$$\text{HS-CF}_\alpha = \frac{\sum_{i=1}^N \min_{\mathbf{p}\in S_\alpha} \|\mathbf{h}_i - \mathbf{p}\|_2}{\sum_{i=1}^N |h_{i,z} - \bar{h}_z| + \epsilon}.$$

The denominator normalizes the nearest-neighbor discrepancy by the depth spread of the visible human points. Lower values indicate that the observed scene surface is closer to the visible body surface.

Let V_h and P_h denote the visible human points and filtered scene points under the human mask. Human–Scene Variance compares their horizontal and vertical spatial extents:

$$\text{HS-V}_\alpha = \frac{1}{2}\Big(|\operatorname{Var}_x(V_h) - \operatorname{Var}_x(P_h)| + |\operatorname{Var}_y(V_h) - \operatorname{Var}_y(P_h)| \Big).$$

HS-V is sensitive to scale and spatial extent, whereas HS-CF is sensitive to surface displacement. Reporting both prevents a translation error and a scale error from being collapsed into one number. Figure 6 illustrates this distinction using matched front and side views. A misplaced body can have approximately the right extent while disagreeing with the scene surface; conversely, a body near the observed person can have the wrong extent. The panels provide a visual interpretation of the two metrics.

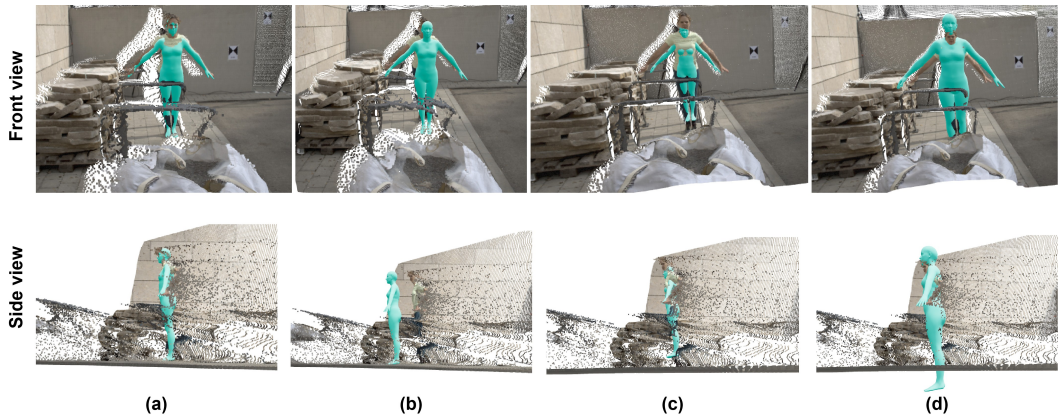


Figure 6: Visual interpretation of human–scene consistency metrics. The front and side views show four body–scene configurations: (a) approximate alignment, (b) positional displacement, (c) a smaller body near the observed location, and (d) a larger body with scale mismatch. HS-CF measures visible-surface proximity, while HS-V compares spatial extent under the human mask.

A.4 Physical Grounding: Table 3

Following the physical-grounding evaluation setting of UniCon3R (Sur et al., 2026), we estimate a robust ground height from the scene and apply a 5 mm tolerance when deciding whether a reconstructed body is in collision with or floating above the support surface. For valid frame f, let $z_f = \min_i v^{\text{up}}_{f,i} - h_s$ be the height of the lowest mesh vertex relative to the estimated ground height h_s, with v^{up} measured along the upward vertical axis. Let $\mathcal{F}$ be valid frames, $\mathcal{F}_{\text{pen}} = \{f \in \mathcal{F} : z_f < -\tau\}$, and $\mathcal{F}_{\text{float}} = \{f \in \mathcal{F} : z_f > \tau\}$, where $\tau = 0.005$ m. Then

$$\text{Coll.} = 100\,\frac{|\mathcal{F}_{\text{pen}}|}{|\mathcal{F}|}, \qquad \text{Pen.} = \frac{100}{|\mathcal{F}_{\text{pen}}|}\sum_{f\in\mathcal{F}_{\text{pen}}}(-z_f),$$

$$\text{Float} = \frac{100}{|\mathcal{F}_{\text{float}}|}\sum_{f\in\mathcal{F}_{\text{float}}} z_f, \qquad \text{P.Max} = 100\max_{f\in\mathcal{F}_{\text{pen}}}(-z_f).$$

Distances are reported in centimeters; an empty event set contributes zero. Lower values indicate more accurate physical grounding.

B Model, Training, and Implementation Details

B.1 Body-Anchored Metric Calibration

Visible SMPL surface samples are projected into the relative depth map. Only finite, positive-depth, in-image samples are retained, and colliding samples keep the nearest human surface. Frames without a detected person or without valid human-depth correspondences are marked invalid and do not contribute to the clip estimate. If the entire clip has no valid correspondences, inference uses unit scale as an explicit fallback and flags the calibration as invalid; that fallback is not counted as a valid metric-scale estimate. The analytic initialization is

$$s^c_t = \underset{(q,i)\in\mathcal{A}_t}{\text{median}} \frac{[\mathbf{V}^{c,(0)}_{tqi}]_z}{D^r_t\left(\pi(\mathbf{K}_t, \mathbf{V}^{c,(0)}_{tqi})\right)}. \tag{10}$$

Twenty-four body anchors encode body state, anchor position and projection, local scene geometry, surface normals, offsets, distances, and depth residuals. The anchor tokens use self-attention over the body and cross-attention to a 3×3 scene neighborhood. The pooled representation predicts residual log-scale and depth bias. Confidence-weighted person predictions are combined per frame, and the valid frame estimates are aggregated in log-scale over the clip before being propagated to scene depth and camera translation.

For metric supervision, let $s^\star$ and $b^\star$ denote the target scene scale and depth bias, and let $\widehat{D}^m$ be the calibrated depth. The calibration objective is

$$\begin{aligned}\mathcal{L}_{\text{calib}} = \lambda_s |\log \widehat{s} - \log s^\star| + \lambda_b |\widehat{b} - b^\star| \\ + \frac{\lambda_d}{|\Omega|} \sum_{\mathbf{u}\in\Omega} |\widehat{D}^m(\mathbf{u}) - D^\star(\mathbf{u})|.\end{aligned} \tag{11}$$

Here Ω contains valid supervised depth pixels. The bias target is zero for the scale-perturbation training path.

B.2 Implicit Interaction Grounding

The SMPL surface is partitioned into 96 corresponding regions; each region is represented by its centroid and eight sampled vertices. In the image, the centroid anchors a fixed 3×3 window and a fixed 7×7 window. The adaptive support follows the projected extent of the region, while the annulus collects context outside that local support. Thus the four supports provide pointwise local evidence, a broader fixed neighborhood, region-sized context, and surrounding scene context.

For each projected region, an instance-aware depth buffer stores the current person's visible SMPL depth. Scene samples that agree with this depth under the depth-consistency tolerance are assigned to the person-owned channel; samples consistent with the observed scene but not owned by this person enter the environment channel. Samples assigned to another person are excluded from both attention channels and summarized as interference features. The tolerance is denoted ϵ_{own}. Attention is masked by valid samples, producing one pooled token per support and ownership channel. An empty support returns a zero token and zero valid-support ratio.

Each region predicts a camera-space translation proposal, a validity gate, and a scalar log-variance. Proposals are aggregated by

$$\begin{aligned}\alpha_r &= m_r g_r \exp(-\ell_r), \\ \widehat{\mathbf{v}} &= \frac{\sum_r \alpha_r \mathbf{v}_r}{\max(\sum_r \alpha_r, \varepsilon)}.\end{aligned} \tag{12}$$

Here m_r is the valid-support indicator, $g_r = \sigma(a_r)$ is the regional gate, and ℓ_r is the predicted log-variance. A person-level gate $g_{\text{p}} = \sigma(a_{\text{p}})$ controls whether the aggregate is applied. The bounded update is

$$\begin{aligned}\Delta\boldsymbol{\tau}^{(k)} &= g_{\text{p}}^{(k)} \,\text{clip}\Big(\widehat{\mathbf{v}}^{(k)}, -\boldsymbol{\delta}_{\max}, \boldsymbol{\delta}_{\max}\Big), \\ \boldsymbol{\tau}^{(k+1)} &= \boldsymbol{\tau}^{(k)} + \Delta\boldsymbol{\tau}^{(k)}.\end{aligned} \tag{13}$$

Here clipping is component-wise. Two refinement iterations share parameters and recompute region locations, ownership, and support after each root-translation update. Pose, shape, camera rotation, and scene geometry remain fixed.

B.3 Training Stages and Freezing Strategy

The scene estimator and human observer are frozen throughout training. We first train on BEDLAM (Black et al., 2023), where metric ground-truth scene depth and SMPL state provide the clean target. Synthetic multiplicative depth-scale perturbations train the calibration module to recover metric scale; residual-head training starts from the analytic scale initialization and learns the remaining scale and depth-bias correction. For interaction grounding, we perturb the ground-truth human root translation and scene scale, including clean, isolated, and mixed states, and train the model to restore the original ground-truth translation. This gives explicit supervision for both recovery from corrupted states and preservation of already-correct inputs. We then fine-tune the grounding module using EMDB (Kaufmann et al., 2023) and 3DPW (Von Marcard et al., 2018) data under the same frozen-backbone setup, using the available human and scene supervision for each dataset.

The grounding objective is

$$\mathcal{L}_{\text{ground}} = \lambda_v \mathcal{L}_{\text{vote}} + \lambda_t \mathcal{L}_{\text{trans}} + \lambda_g \mathcal{L}_{\text{gate}} + \lambda_r \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{state}}. \tag{14}$$

At iteration k, let $\boldsymbol{\tau}^\star$ be the target root translation and K the number of remaining updates. The vote target is the remaining displacement $\mathbf{d}^{(k)} = (\boldsymbol{\tau}^\star - \boldsymbol{\tau}^{(k)})/K$; $\mathcal{L}_{\text{vote}}$ is the support- and reliability-weighted mean absolute error between each valid regional proposal and $\mathbf{d}^{(k)}$. $\mathcal{L}_{\text{trans}}$ is the mean

absolute error between the final refined root translation and $\boldsymbol{\tau}^\star$. The gate target is one when applying the candidate update improves root-translation error and zero otherwise, and $\mathcal{L}_{\text{gate}}$ is binary cross-entropy on valid examples. $\mathcal{L}_{\text{reg}}$ penalizes predicted update magnitude. $\mathcal{L}_{\text{state}}$ combines a recovery term for perturbed states, an identity term that discourages movement of clean inputs, and a no-worse term that penalizes updates that increase translation error. More explicitly, let $\mathcal{P}$ be the perturbed training examples and $\mathcal{C}$ the clean examples. The state term is

$$\mathcal{L}_{\text{state}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{nw}}\mathcal{L}_{\text{nw}}. \tag{15}$$

Here $\mathcal{L}_{\text{rec}}$ is the mean L_1 recovery error on perturbed examples, $\mathcal{L}_{\text{id}}$ is the mean L_1 change on clean inputs, and $\mathcal{L}_{\text{nw}}$ penalizes any increase in translation error. Regional votes are supervised at each iteration against the remaining displacement; the final translation term supervises the last iterate; the binary gate target is positive when the candidate update reduces root-translation error; and the update regularizer is the mean L_1 norm of the applied translation increments.

B.4 Inference Algorithm

Algorithm S1: GroundedHuman inference

Input: $\{\mathbf{I}_t\}_{t=1}^{T}$, frozen f_Ω and f_{NLF}, trained f_{cal} and f_{grd}.
Output: Metric scene points $\{\mathbf{P}_t^w\}$ and meshes $\{\mathbf{M}_{tq}^w\}$.

$(D^r, \mathbf{K}, \mathbf{T}^r, \mathbf{F}) \leftarrow f_\Omega(\{\mathbf{I}_t\}_{t=1}^{T})$
$\{(\boldsymbol{\theta}_{tq}, \boldsymbol{\beta}_{tq}, \boldsymbol{\tau}_{tq}^{(0)}, c_{tq})\} \leftarrow f_{\text{NLF}}(\{\mathbf{I}_t\}_{t=1}^{T})$
$\mathcal{A}_t \leftarrow \text{ProjectVisibleSMPL}(D_t^r, \mathbf{K}_t, \boldsymbol{\theta}_{tq}, \boldsymbol{\beta}_{tq})$
$s_t^{(0)} \leftarrow \text{median}_{(q,i)\in\mathcal{A}_t} \dfrac{[\mathbf{V}_{tqi}^{c,(0)}]_z}{D_t^r(\pi(\mathbf{K}_t, \mathbf{V}_{tqi}^{c,(0)}))}$
$(\Delta \log s_t, b_t) \leftarrow f_{\text{cal}}(\mathcal{A}_t, \mathbf{F}_t, D_t^r)$
$(s, b) \leftarrow \text{AggregateValid}_t(s_t^{(0)} \exp(\Delta \log s_t), b_t)$
$D_t^m \leftarrow sD_t^r + b, \qquad \mathbf{t}_{cw,t}^m \leftarrow s\mathbf{t}_{cw,t}^r$
$\boldsymbol{\tau}_{tq}^{(0)} \leftarrow \text{MetricInit}(\boldsymbol{\tau}_{tq}^{(0)}, s, b)$
for $k = 0, 1$ **do**
  $\mathbf{R}_{tq}^{(k)} \leftarrow \text{Partition}_{96}(\mathbf{M}_{tq}(\boldsymbol{\tau}_{tq}^{(k)}))$
  $\mathbf{x}_r^{(k)} \leftarrow \text{Pool}_{\mathcal{S}_1,\ldots,\mathcal{S}_4}(\mathbf{F}, D^m, \mathbf{R}_{tq}^{(k)}), \quad r \in \mathbf{R}_{tq}^{(k)}$
  $(\mathbf{v}_r^{(k)}, g_r^{(k)}, \ell_r^{(k)}) \leftarrow f_{\text{grd}}(\mathbf{x}_r^{(k)})$
  $\alpha_r^{(k)} \leftarrow m_r^{(k)} g_r^{(k)} e^{-\ell_r^{(k)}}, \qquad \widehat{\mathbf{v}}^{(k)} \leftarrow \dfrac{\sum_r \alpha_r^{(k)} \mathbf{v}_r^{(k)}}{\max(\sum_r \alpha_r^{(k)}, \varepsilon)}$
  $\boldsymbol{\tau}_{tq}^{(k+1)} \leftarrow \boldsymbol{\tau}_{tq}^{(k)} + g_{\text{p}}^{(k)} \,\text{clip}(\widehat{\mathbf{v}}^{(k)}, -\boldsymbol{\delta}_{\max}, \boldsymbol{\delta}_{\max})$
end for
$\mathbf{P}_t^w, \mathbf{M}_{tq}^w \leftarrow \text{TransformToWorld}(D_t^m, \mathbf{T}_t^m, \boldsymbol{\tau}_{tq}^{(2)}, \boldsymbol{\theta}_{tq}, \boldsymbol{\beta}_{tq})$

Region and support features are recomputed at each iteration.

B.5 Runtime Measurement

The steady-state runtime is measured on an NVIDIA A800-SXM4-80GB using a 100-frame input tensor of shape $[1, 100, 3, 448, 592]$, resized to a 512-pixel balanced input. After five warm-up iterations, 20 timed iterations are run with device synchronization. Image I/O, resizing, host-to-device transfer, model loading, and warm-up are excluded. The frame rate is computed from the number of input frames and the mean synchronized latency:

$$\text{FPS} = \frac{N_{\text{frames}}}{\frac{1}{N_{\text{iter}}}\sum_{j=1}^{N_{\text{iter}}} t_j},$$

where t_j is the synchronized wall-clock latency of timed iteration j.

C Supplementary Results and Analysis

C.1 Ablation Reading Guide

Table 3 removes one grounding component at a time while retaining metric calibration and the frozen pose and shape. The metrics are defined in Sec. A.4. Disabling grounding gives the largest collision ratio, mean penetration, and floating distance. The complete model gives the lowest collision ratio and maximum penetration. The component ablations are not uniformly worse on every metric: removing multi-scale context gives the lowest mean penetration, while removing reliability weighting gives the lowest floating distance but the highest maximum penetration.

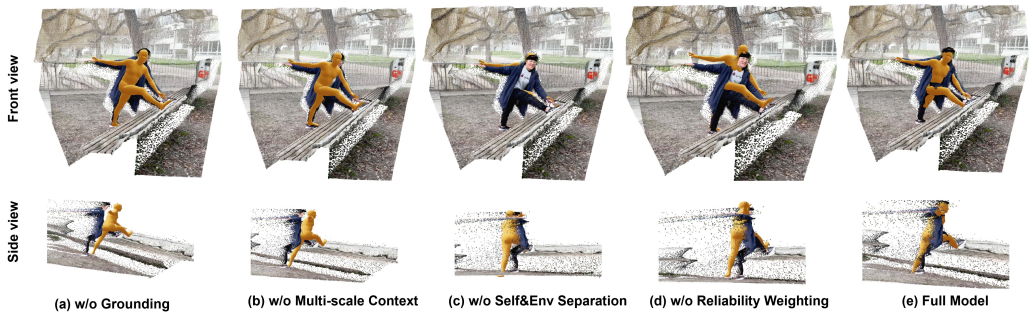


Figure 7: Visual ablation of Implicit Interaction Grounding. Matched front and side views compare (a) no grounding, (b) no multi-scale context, (c) no self/environment separation, (d) no reliability weighting, and (e) the full model.

D Qualitative Reconstructions

D.1 Long-Sequence Reconstructions

Figure 8 presents sampled body meshes within reconstructed scene geometry across longer sequences. The upper and lower-right panels show repeated body poses along paths through outdoor scenes; the lower-left panel shows varied articulated poses near scene surfaces. Together, the panels show temporal progression, scene-relative placement, and articulation.

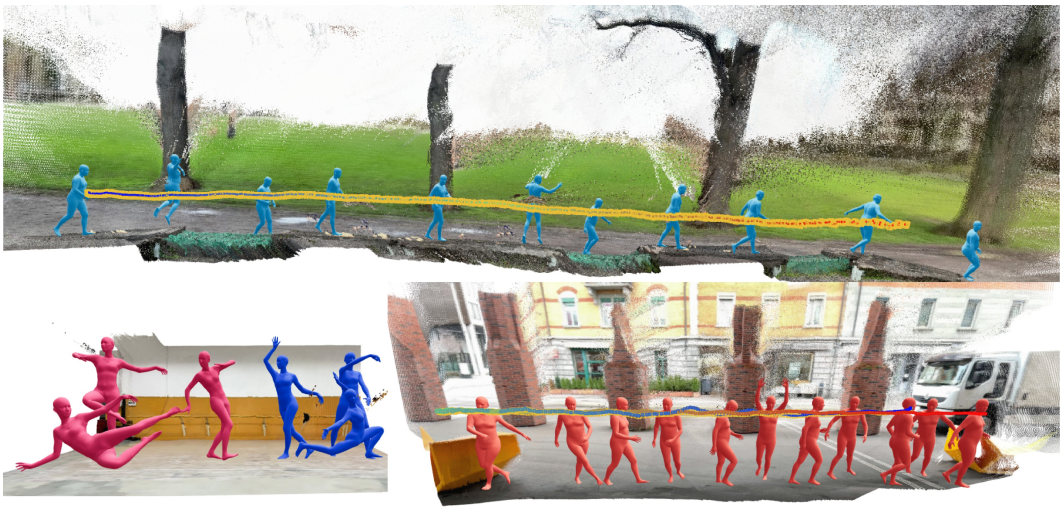


Figure 8: Qualitative montage of human and scene reconstruction over multiple frames. Top: sampled human meshes and an overlaid path in an outdoor scene. Bottom left: varied articulated poses placed near scene geometry. Bottom right: sampled poses and a path across a second street scene.

D.2 Metric Measurements on PROX

On PROX (Hassan et al., 2019), we compare reconstructed human height and furniture dimensions with reference measurements and MapAnything (Keetha et al., 2026) reconstructions (Fig. 9). The standing-person example places body height and table length in the same coordinate system, while the seated example relates the person to the surrounding partitions. These measurements complement the placement comparison: a body may appear aligned with nearby surfaces from one viewpoint even when the overall scene dimensions are inaccurate. Inspecting both human and environmental lengths therefore provides a physical interpretation of the joint geometry beyond its projected appearance. The examples also expose how scale affects objects at different positions within the scene, rather than only the human used as a reference.
Together, these evaluations connect local body geometry, global motion, and scene-relative placement within a single reconstruction.

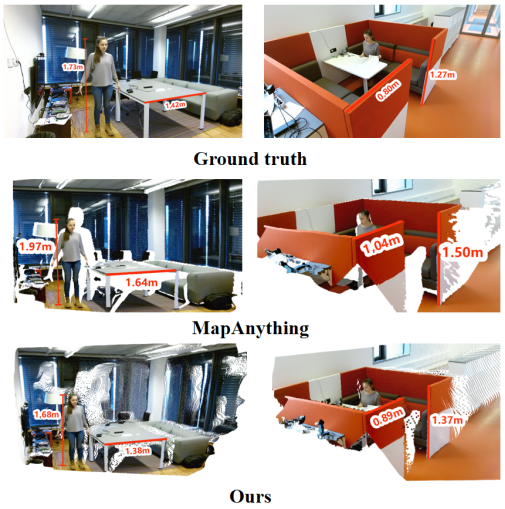


Figure 9: PROX metric measurements. Rows show ground truth, MapAnything, and GroundedHuman; columns show standing and seated examples.

D.3 Foot–Ground Plausibility Comparison

Figure 10 compares foot–ground plausibility across GroundedHuman, UniSH (Li et al., 2026), Human3R (Chen et al., 2025), and UniCon3R (Sur et al., 2026) on the same input frame. The dashed regions highlight differences in foot placement relative to the reconstructed ground.

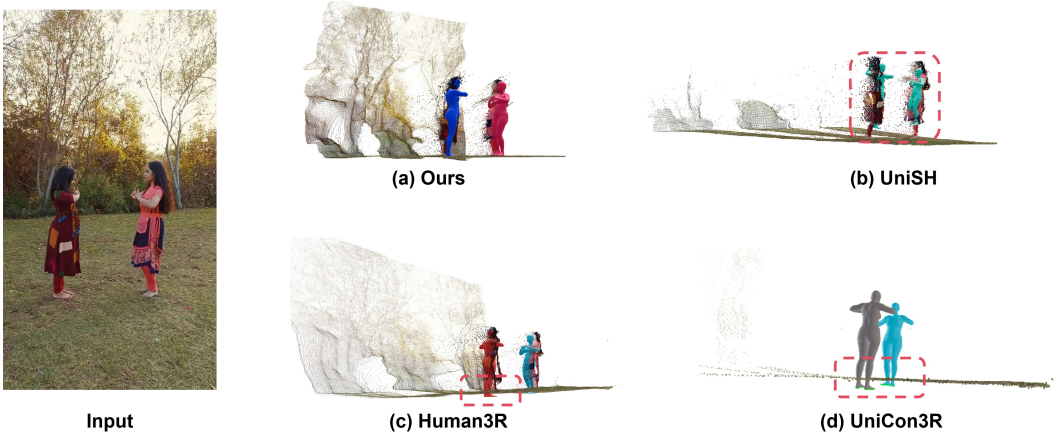


Figure 10: Foot–ground plausibility comparison across GroundedHuman, UniSH, Human3R, and UniCon3R.The input image is shown at left; GroundedHuman, UniSH, Human3R, and UniCon3R reconstructions are shown at right. Reddashed boxes highlight regions for inspecting foot placement relative to the reconstructed ground.