UniCon3R: Unified Contact-aware 4D Human-Scene Reconstruction from Monocular Video

Tanuj Sur[1] **Shashank Tripathi**[2] **Nikos Athanasiou**[2]
Ha Linh Nguyen[1] **Kai Xu**[1] **Michael J. Black**[2] **Angela Yao**[1]
[1]National University of Singapore
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
{tanujsur,halinh,kxu,ayao}@comp.nus.edu.sg
{stripathi,nathanasiou,black}@tuebingen.mpg.de

Abstract

We introduce **UniCon3R**, a unified feed-forward framework for online human-scene 4D reconstruction from monocular video. Current feed-forward human-scene reconstruction methods suffer from artifacts, where bodies float above the ground or penetrate parts of the scene. A key reason is the lack of effective interaction modelling between the human and the environment. Our goal is to exploit contact between the human and the scene during inference to actively improve the human mesh reconstruction. To that end, we explicitly model interaction by inferring 4D contact from the human pose and scene geometry and use the contact as a corrective cue for generating the pose. This enables UniCon3R to jointly recover scene geometry and spatially aligned 4D humans within the scene. Experiments on standard human-centric video benchmarks show that UniCon3R outperforms state-of-the-art baselines on physical plausibility and global human motion estimation while preserving fast, feed-forward inference speeds. The results validate our central claim: contact serves as a powerful internal prior, thus establishing a new paradigm for physically grounded joint human-scene reconstruction. Project page: https://surtantheta.github.io/UniCon3R.

1 Introduction

Recent work on 3D and 4D human-scene reconstruction [7, 20, 22, 26, 45, 48–50, 53] from monocular images and videos can recover human meshes and their surrounding environments. Unified feed-forward methods [7, 20, 53] are especially promising because they predict global human motion, dense scene geometry, and camera trajectories within a single forward pass, enabling online 4D reconstruction. However, joint scene reconstruction and Human Mesh Recovery (HMR) [2, 5, 10, 12, 18, 19, 33, 36] in a world coordinate frame does not by itself guarantee physical plausibility [37, 39, 48]. Specifically, a body may be accurately localized in 3D yet still float above the floor, penetrate objects, or slide over supporting surfaces. In everyday human-scene interaction, physical coupling with the environment is mediated by contact: the body is supported by scene surfaces and interacts with objects through local contact regions. We posit that the 3D contacts between the human and the scene provide important constraints that could improve the joint reconstruction.

Unfortunately, contact estimation is inherently ill-posed because the contact areas are often occluded in 2D images [8, 11, 35]. Existing unified feed-forward methods [7, 20, 22] handle the physical interaction between the human and the scene implicitly, as the human reconstruction branch does not leverage scene-grounded interaction cues during inference. Resulting reconstructions exhibit artifacts such as floating feet, penetrations, or inconsistent dynamics, as shown in Fig. 1(a). This leaves a clear gap between merely reconstructing a person in a scene and physically plausible reconstructions.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

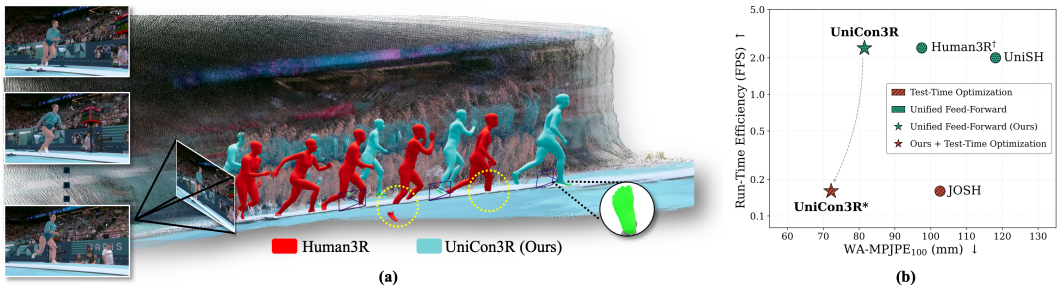

Figure 1: **Contact improves unified 4D human-scene reconstruction. (a)** Human3R [7] meshes are misaligned, with the feet penetrating the ground (dashed yellow circles). UniCon3R predicts dense per-vertex contact (green inset) and uses it to improve body-scene alignment. **(b) Amortized runtime versus global human motion error** . UniCon3R achieves the state-of-the-art on the RICH dataset [14]; test-time optimization further reduces the error, but also reduces runtime efficiency.

To address this gap, we introduce **UniCon3R**, a contact-aware framework for unified feed-forward [1] 4D human-scene reconstruction. Our key insight is to treat contact as an internal feedback signal that interacts with both the human and scene reconstruction branches, enabling more accurate human localization within the forward pass. Building upon the prompting mechanisms of CUT3R [42] and Human3R [7], we introduce a *scene-aware contact prompt* that gathers physical cues from the current frame, recurrent scene memory, and local scene geometry. The resulting contact prompt is decoded jointly with the human and scene tokens, allowing contact cues to influence the latent representation used for final human mesh reconstruction.

In principle, a model that predicts contact should help improve the body's world-frame placement [8, 48]. However, we find that this alone is not sufficient. As our experiments show, such a model may correctly identify where contact is likely to occur, yet still fail to localize the body accurately. Thus, the key challenge is to ensure that predicted contact actively shapes the recovered human mesh within the unified feed-forward pipeline [47, 48].

To this end, rather than using contact as an auxiliary prediction [11, 35], we feed the inferred contact cues back into the HMR branch via a *contact-guided latent refinement*. This allows the refined contact token to produce a residual update to the human latent representation during the forward pass. By internalizing contact, UniCon3R preserves the simplicity and streaming capabilities of unified feed-forward inference Fig. 1(b) while ensuring the recovered human mesh is physically plausible in scene-afforded constraints. Notably, incorporating contact into 4D human-scene reconstruction also boosts contact estimation accuracy, highlighting the strong synergy between the two tasks.

We evaluate UniCon3R on the challenging human-centric video benchmarks RICH [14], EMDB [17], 3DPW [41] and SLOPER4D [9], focusing on physical plausibility, global human motion estimation and local human mesh recovery. Our results show that internalizing contact as a refinement signal reduces penetration by **74%**, and improves WA-MPJPE metric by **16%** on the RICH [14] dataset, while preserving online feed-forward inference. In addition, on RICH [14], UniCon3R improves contact recall over DECO [35] while achieving higher F1 and lower geometric contact error.

In summary, our contributions are as follows: (i) We introduce a new design paradigm for unified feed-forward 4D human-scene reconstruction by modeling human-scene contact as an *internal corrective signal*. (ii) We propose **UniCon3R**, a novel contact-aware framework that internalizes human-scene interaction within unified feed-forward reconstruction through a *scene-aware contact prompt* and *contact-guided latent refinement*. (iii) We show that actively internalizing contact improves physical plausibility, global human motion and local human mesh recovery over existing state-of-the-art feed-forward methods, while preserving competitive streaming efficiency.

2 Related Work

3D Human Reconstruction from monocular views has evolved from optimization-based fitting to feed-forward regression. Early methods such as SMPLify [4] fit parametric body models such as

[1] Following CUT3R [42], we use *feed-forward* to mean direct online inference without per-scene optimization or global alignment, even though the model maintains a recurrent persistent state.

SMPL [23] or SMPL-X [28] to image evidence. More recent HMR methods directly regress body parameters from images or videos [5, 10, 12, 16, 21, 27, 36, 43, 46, 52], improving robustness and accuracy under challenging viewpoints, occlusions, and motion. Some recent methods also estimate global human motion in world coordinates from monocular video, but typically without explicitly modeling physical interaction with the surrounding 3D scene. In contrast, our goal is not only to recover a plausible human body, but to refine it within a jointly reconstructed metric scene, where scene-grounded contact acts as an internal corrective cue for the final human estimate.

3D Human-Scene Contact provides direct evidence of how the body is supported by the environment. RICH [14] introduced dense full-body human-scene contact labels by combining multiview human capture with scanned 3D scenes, and proposed BSTRO for vertex-level body-scene contact estimation from images. DECO [35] predicts dense 3D contact on the SMPL body from in-the-wild images using body-part and scene-context attention. InteractVLM [11] estimates contact on both humans and objects by lifting 2D foundation-model predictions into 3D. PICO [8] further studies dense body-object contact correspondences and uses them for optimization-based human-object reconstruction. However, contact is typically treated either as a separate prediction target or as a signal used only in a later fitting stage. For example, JOSH [22] uses BSTRO [14] as an off-the-shelf contact model to refine body pose during motion estimation. In contrast, we incorporate contact directly into the feed-forward reconstruction pathway.

Joint Human-Scene Reconstruction recovers people together with the surrounding 3D environment. Early methods often rely on optimization and are restricted to static or multi-view settings [26, 29, 30]. Later monocular methods couple human priors with scene geometry. SyncHMR [54] uses SMPL-based constraints for metric-scale reconstruction, while JOSH [22] initializes from dense scene reconstruction and human mesh recovery, then optimizes scene geometry, camera poses, and human motion with contact constraints. These methods can improve body-scene consistency, but require expensive per-scene optimizations.

Recent methods reduce this cost with feed-forward or optimization-free designs. JOSH3R [22] distills JOSH pseudo-labels into an optimization-free model. Human3R [7] adapts CUT3R [42] with visual prompt tuning to predict humans, scenes, and camera motion in a unified feed-forward model while keeping the scene backbone frozen. UniSH [20] combines scene-reconstruction and HMR priors to recover scene geometry, camera parameters, and metric-scale SMPL bodies in a single forward pass. These methods speed up joint reconstruction but human-scene interaction is still implicit, which often leads to physically implausible results. Our method models contact explicitly and feeds the contact representation back into the HMR pathway before the body regression.

3 Preliminaries

World-Grounded Motion & 4D Scene Reconstruction. Modeling human-scene interaction requires representing the human and camera in a common world coordinate frame. In this work, we adopt the SMPL-X [28] human model and express the world-frame body vertices as $\mathbf{V}_t^{\text{body}} \in \mathbb{R}^{V\times 3}$, with $\mathbf{V}_t^{\text{body}} = \text{SMPL-X}(\boldsymbol{\theta}_t, \boldsymbol{\beta}, \boldsymbol{\alpha}_t, \mathbf{P}_t)$. Here, $\boldsymbol{\theta}_t$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}_t$ denote the pose, shape, and expression parameters of the SMPL-X model. $\mathbf{P}_t^{\text{cam}}$ is the local root translation in the camera frame and $\mathbf{T}_t$ is the camera pose that maps this local translation into the world frame.

Our scene reconstruction module builds on CUT3R's [42] 4D scene reconstruction model. As shown in Fig. 2 (top), given an image stream, it maintains a latent scene state $\mathbf{S}_t$ updated at each frame. Here, t denotes the frame index. At time step t, the input image is encoded into feature tokens $\mathbf{F}_t$, which interact with the previous state $\mathbf{S}_{t-1}$ together with a learnable pose token $\mathbf{z} \in \mathbb{R}^c$, where c is the decoder embedding dimension. This produces updated tokens $\mathbf{F}'_t$, $\mathbf{z}'_t$, and an updated state $\mathbf{S}_t$. Two separate heads decode camera-frame and world-frame scene pointmaps, $\mathbf{X}_t^{\text{cam}}$ and $\mathbf{X}_t^{\text{world}}$, and regresses the camera pose $\mathbf{T}_t$ from the updated tokens $\mathbf{F}'_t$, $\mathbf{z}'_t$.

4D Human Reconstruction via visual prompting. For reconstructing the human mesh, we follow Human3R [7], which extends CUT3R [42] from pure scene reconstruction to joint human-scene reconstruction through parameter-efficient Visual Prompt Tuning (VPT) [15]. Rather than modifying the frozen CUT3R backbone, Human3R adds a human prompt $\mathbf{H}_t \in \mathbb{R}^{N_t\times c}$ where N_t denotes the number of detected humans in frame t. As shown in Fig. 2 (middle), we detect human anchors

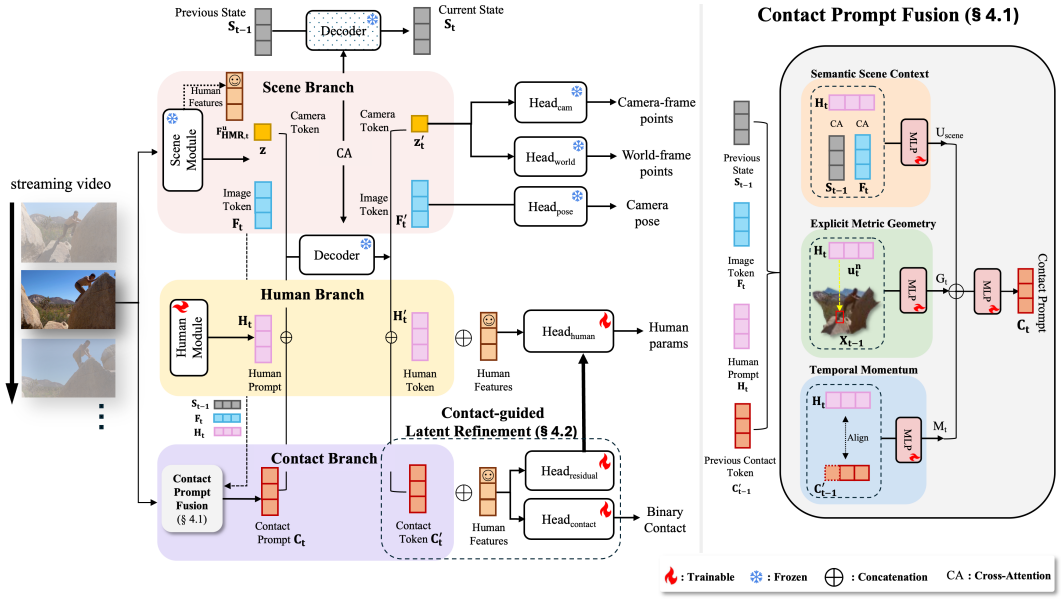

Figure 2: **Overview of UniCon3R. Left:** At time t, the Scene and Human branch encodes the current frame into image tokens $\mathbf{F}_t$ and a human prompt $\mathbf{H}_t$, to interact with the previous state $\mathbf{S}_{t-1}$ and update the current state $\mathbf{S}_t$, to predict scene, camera, and human parameters. Our contact branch outputs dense vertex-level contact on the SMPL [23] mesh. **Right:** UniCon3R constructs a *scene-aware contact prompt* $\mathbf{C}_t$ (§4.1) from semantic scene context, explicit metric geometry, and a temporal momentum term. This gets decoded into a refined contact token $\mathbf{C}'_t$ which predicts dense SMPL contact. It also drives *contact-guided latent refinement* (§4.2) by producing a residual update to the human latent before final SMPL-X regression.

$\{\mathbf{u}_t^n\}_{n=1}^{N_t}$ from the CUT3R image tokens and gather the corresponding human-prior features $\mathbf{F}^{\mathbf{u}}_{\mathrm{HMR},t}$ from a frozen Multi-HMR encoder [2].

The human prompt in the HMR pathway acts as a human-specific query: it self-attends to image tokens to aggregate body evidence and cross-attends to the persistent scene state to retrieve temporally consistent human information anchored in the world-centered scene context. The interaction can be summarized as: $[\mathbf{F}'_t, \mathbf{z}'_t, \mathbf{H}'_t], \mathbf{S}_t = \mathrm{Decoders}([\mathbf{F}_t, \mathbf{z}, \mathbf{H}_t], \mathbf{S}_{t-1})$, $\mathbf{Y}_t = \mathrm{Head}_{\mathrm{human}}(\mathbf{H}'_t)$ where $\mathbf{Y}_t$ denotes the predicted SMPL-X [28] parameters. Here, blue denotes frozen modules inherited from CUT3R and Human3R, while red denotes learnable modules introduced or fine-tuned in our framework. These preliminaries expose the key limitation of Human3R-style unified reconstruction: the interaction between $\mathbf{F}'_t$ and $\mathbf{H}_t$ cannot ensure physically grounded reconstruction.

4 UniCon3R

UniCon3R is a contact-aware framework for online 4D human-scene reconstruction. Contact is encoded as an explicit prompt, processed by the same frozen decoder-and-memory pipeline used for scene and human reconstruction, and then used to refine the recovered body. The method has two components: **(1)** a scene-aware contact prompt that aggregates interaction evidence from the current frame, recurrent scene memory, and local scene geometry, and **(2)** a contact-guided latent refinement mechanism that feeds the refined contact representation back into the HMR pathway before final SMPL-X regression [28]. Fig. 2 shows an overview of the framework.

4.1 Contact Prompt Fusion

Semantic Scene Context. Robust contact reasoning requires immediate spatial evidence (e.g., observing a shoe hitting the floor). To recover such cues, we extract two complementary scene features, $\mathbf{U}_{\mathrm{curr}} = \mathrm{CA}_{\mathrm{curr}}(\mathbf{H}_t, \mathbf{F}_t)$, $\mathbf{U}_{\mathrm{mem}} = \mathrm{CA}_{\mathrm{mem}}(\mathbf{H}_t, \mathbf{S}_{t-1})$ where $\mathrm{CA}(\cdot,\cdot)$ denotes cross-attention. Here, $\mathbf{U}_{\mathrm{curr}}$ captures visual evidence from the current frame, while $\mathbf{U}_{\mathrm{mem}}$ retrieves persistent object context from the recurrent scene state. We combine these cues through a learned

gate: $\boldsymbol{\gamma}_t = \sigma(\mathrm{MLP}(\mathbf{H}_t \oplus \mathbf{U}_{\mathrm{curr}} \oplus \mathbf{U}_{\mathrm{mem}}))$, $\mathbf{U}_{\mathrm{scene}} = \boldsymbol{\gamma}_t \odot \mathbf{U}_{\mathrm{curr}} + (1 - \boldsymbol{\gamma}_t) \odot \mathbf{U}_{\mathrm{mem}}$ where $\odot$ is the element-wise multiplication, $\sigma(\cdot)$ denotes the element-wise sigmoid function, so that $\boldsymbol{\gamma}_t \in [0, 1]^c$ acts as a soft gate for balancing current-frame and previous scene context.

Explicit Metric Geometry. Scene context features alone do not encode metric distances between interacting parts explicitly, so we inject local 3D support cues through a geometry token. For notational convenience, we write $\mathbf{X}_{t-1} := \mathbf{X}_{t-1}^{\mathrm{world}}$ for the previous-frame world-frame pointmap. Let $\mathbf{u}_t^n \in \mathbb{R}^2$ denote the current anchor of person n. For each detected human, we encode the local geometry around $\mathbf{u}_t^n$ into the prompt space:

$$\mathbf{G}_t = \{\mathrm{MLP}(\phi_{\mathrm{geo}}(\mathbf{X}_{t-1}, \mathbf{u}_t^n))\}_{n=1}^{N_t} \in \mathbb{R}^{N_t \times c}, \tag{1}$$

where $\phi_{\mathrm{geo}}(\mathbf{X}_{t-1}, \mathbf{u}_t^n) \in \mathbb{R}^3$ pools a local pointmap neighborhood centered at $\mathbf{u}_t^n$. Concretely, we construct a fixed-size square window centered at $\mathbf{u}_t^n$ and apply RoIAlign [13] to $\mathbf{X}_{t-1}$.

Temporal Momentum. Although realistic 3D contact is temporally coherent, reasoning contact from images often tends to be unstable due to strong challenges such as occlusions and depth-scale ambiguities. To stabilize interaction reasoning over time, we reuse the contact token from the previous frame $\mathbf{C}'_{t-1}$ and project it into the current contact prompt latent space as:

$$\mathbf{M}_t = \mathrm{MLP}\big(\mathrm{Align}\big(\mathbf{C}'_{t-1}; N_t\big)\big) \in \mathbb{R}^{N_t \times c}, \tag{2}$$

where $\mathrm{Align}(\cdot; N_t)$ keeps the first $\min(N_{t-1}, N_t)$ rows of $\mathbf{C}'_{t-1}$, truncates extra rows when $N_{t-1} > N_t$, and appends zero rows when $N_{t-1} < N_t$.

Fusion. We fuse human, semantic, geometric, and temporal cues to obtain the final contact token $\mathbf{C}_t = \mathrm{MLP}(\mathbf{H}_t \oplus \mathbf{U}_{\mathrm{scene}} \oplus \mathbf{G}_t \oplus \mathbf{M}_t)$ as shown in Fig. 2 (right).

4.2 Contact-Guided Latent Refinement

A contact prompt is only useful if it can influence the final body estimate. Unlike Parallel Readout, which predicts contact as an auxiliary output, we use the refined contact prompt $\mathbf{C}'_t$ to update the human latent before final SMPL-X regression as shown in Fig. 2 (bottom). First, the human and the contact prompts $\mathbf{H}_t$ and $\mathbf{C}_t$ are processed together by the frozen 4D decoders (§ Section 3):

$$[\mathbf{F}'_t, \mathbf{z}'_t, \mathbf{H}'_t, \mathbf{C}'_t], \mathbf{S}_t = \mathrm{Decoders}([\mathbf{F}_t, \mathbf{z}_t, \mathbf{H}_t, \mathbf{C}_t],\ \mathbf{S}_{t-1})\,. \tag{3}$$

After interaction with the previous state $\mathbf{S}_{t-1}$, we augment $\mathbf{H}'_t$ and $\mathbf{C}'_t$ with the human prior features $\mathbf{F}^{\mathbf{u}}_{\mathrm{HMR},t}$ obtained from the Multi-HMR encoder as

$$\tilde{\mathbf{H}}_t = \mathbf{H}'_t \oplus \mathbf{F}^{\mathbf{u}}_{\mathrm{HMR},t}, \qquad \tilde{\mathbf{C}}_t = \mathbf{C}'_t \oplus \mathbf{F}^{\mathbf{u}}_{\mathrm{HMR},t}. \tag{4}$$

The refined contact token $\tilde{\mathbf{C}}_t$ is decoded to obtain dense per-vertex contact logits $\mathbf{s}_t^v$ and probabilities $\hat{\mathbf{c}}_t^v$ on the SMPL mesh ($V_s = 6{,}890$ vertices):

$$\hat{\mathbf{c}}_t^v = \sigma(\mathbf{s}_t^v)\,, \quad \text{where } \mathbf{s}_t^v = \mathrm{Head}_{\mathrm{contact}}\Big(\tilde{\mathbf{C}}_t\Big) \in \mathbb{R}^{N_t \times 6890}. \tag{5}$$

The final body mesh $\mathbf{Y}_t$ is regressed from a human latent updated by a residual $\Delta\mathbf{H}_t$ based on $\tilde{\mathbf{C}}_t$:

$$\mathbf{Y}_t = \mathrm{Head}_{\mathrm{human}}\big(\bar{\mathbf{H}}_t\big)\,, \quad \text{where } \bar{\mathbf{H}}_t = \tilde{\mathbf{H}}_t + \Delta\mathbf{H}_t \text{ and } \Delta\mathbf{H}_t = \mathrm{Head}_{\mathrm{residual}}\Big(\tilde{\mathbf{C}}_t\Big)\,. \tag{6}$$

Training Objectives. The overall objective of UniCon3R $\mathcal{L}_{\mathrm{total}}$ is the sum of $\mathcal{L}_{\mathrm{4D}}, \mathcal{L}_{\mathrm{SMPLX}}, \mathcal{L}_{\mathrm{contact}}$. Here, $\mathcal{L}_{\mathrm{4D}}$ denotes the 4D reconstruction losses for metric pointmaps, camera pose, and appearance consistency inherited from CUT3R [42], while $\mathcal{L}_{\mathrm{SMPLX}}$ supervises the human output $\mathbf{Y}_t$ using standard SMPL-X [28] parameter, mesh, joint, and reprojection losses [7, 42].

The contact loss is $\mathcal{L}_{\mathrm{contact}} = \mathcal{L}_c^{3D} + \lambda_{\mathrm{p}}\mathcal{L}_p$, with dense vertex-level supervision $\mathcal{L}_c^{3D} = \mathrm{FocalBCE}\big(\mathbf{s}_t^v, \mathbf{c}_{\mathrm{gt}}^v\big)$. Here, $\mathbf{s}_t^v \in \mathbb{R}^{N_t \times 6890}$ are predicted SMPL [23] contact logits, $\mathbf{c}_{\mathrm{gt}}^v$ are binary vertex-level contact labels, and $\mathcal{L}_p$ is a part-level contact loss following DECO [35]. Please refer to **Sup. Mat.** (Section B)for training details.

Table 1: **Global human motion estimation on RICH [14] and EMDB-2 [17]. Scene** denotes 3D scene reconstruction, and **Contact** denotes dense human-scene contact prediction or use. Human3R*: released checkpoint in our pipeline; Human3R†: same architecture with our training protocol. WA-MPJPE, W-MPJPE, and PA-MPJPE are reported in mm; RTE is reported in %. **Bold** = best within feed-forward 4D human-scene reconstruction methods; underlined = second-best.

Method	Scene	Contact	RICH [14]				EMDB-2 [17]			
			WA-MPJPE ↓	W-MPJPE ↓	RTE (%) ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE (%) ↓	PA-MPJPE ↓
JOSH3R [22]	✓	✗	190.4	334.9	6.3	–	220.0	661.7	13.1	–
UniSH [20]	✓	✗	118.1	183.2	4.8	50.0	118.5	**270.1**	5.8	44.4
Human3R* [7]	✓	✗	109.9	184.0	3.3	48.6	118.2	286.1	2.4	38.7
Human3R† [7]	✓	✗	97.5	153.2	2.3	41.4	–	–	–	–
Parallel Readout (Ours)	✓	✓	97.9	153.5	2.3	41.5	160.7	388.3	3.6	50.9
UniCon3R (Ours)	✓	✓	**81.5**	**129.8**	**1.7**	**37.6**	**113.7**	285.6	**2.3**	**38.2**

5 Experiments

Datasets. For global human motion estimation, we evaluate on RICH [14] and EMDB-2 [17] following prior works [7, 20, 22, 53]. For local human mesh recovery, we evaluate on SLOPER4D [9] and 3DPW [41]. For physical plausibility and contact prediction, we evaluate on RICH [14], since it provides scene geometry, human motion, and dense body-scene contact annotations in world frame.

Metrics. Following prior work [7, 32, 44], for global human motion estimation we report WA-MPJPE (mm), W-MPJPE (mm), and RTE (%). For local human mesh recovery, we report PVE (mm) and MPJPE (mm). For physical plausibility, we report Collision Ratio (%), Penetrate (cm), Float (cm), and Pen. Max (cm) using the grounding protocol adapted from HuMoS [37]. For contact prediction, we report Precision, Recall, F1, and geometric error, following prior works [11, 14, 35].

Baselines. We primarily compare with feed-forward monocular 4D human-scene reconstruction methods, as they share UniCon3R's setting: jointly predicting humans, scene geometry, and camera motion without per-scene optimisation. In particular, we compare against Human3R [7], UniSH [20], and JOSH3R [22]. We also introduce two controlled baselines for fair architectural comparisons. Human3R† keeps the original Human3R architecture but uses the same training protocol as UniCon3R, in which we fine-tune on BEDLAM [3] and RICH [14] datasets. *Parallel Readout* adds an auxiliary branch as output to the same backbone without a dedicated contact token, and does not feed contact features back into human reconstruction. Throughout the experiments, Human3R* denotes the released Human3R checkpoint evaluated under the same preprocessing, inference, and evaluation pipeline used for UniCon3R.

5.1 Global Human Motion Estimation and Local Human Mesh Recovery

On **global human motion estimation** (see Table 1), UniCon3R improves over all Human3R-derived baselines on both RICH [14] and EMDB-2 [17]. Human3R† controls for gains induced by the training protocol, while *Parallel Readout* shows the gains come from feeding contact into reconstruction, not predicting it in parallel. On EMDB-2, UniCon3R again improves over all Human3R-derived baselines, achieving the best WA-MPJPE, RTE, and PA-MPJPE, and the second-best W-MPJPE overall. These results confirm that UniCon3R's gains generalize beyond RICH. On W-MPJPE, which captures long-horizon trajectory drift by aligning predictions only using the first two frames, UniSH obtains a lower error on EMDB-2, likely benefiting from its sequence-level π^3 reconstruction backbone. In contrast, UniCon3R achieves better local pose accuracy, as reflected by the lower WA-MPJPE, while operating at fast inference speeds (2.4 fps vs 2.0 fps for UniSH), see Fig. 1 (b), in a streaming setting. Please refer to **Sup. Mat.** (Section A.2) for comparisons against offline optimization methods [22, 44, 51] and world-grounded HMR [31, 32, 34, 53] methods. In **Sup. Mat.** (Section A.4), we also report global human motion metrics across different backbones to show that contact-aware UniCon3R's gains are agnostic to the choice of the underlying backbone.

On **local human mesh recovery** (see Table 2), UniCon3R gives the best PVE and MPJPE among baselines on SLOPER4D [9]. On the 3DPW [41] dataset, UniCon3R obtains the lowest MPJPE and PVE comparable to Human3R*. Overall, the results show that UniCon3R improves global human motion estimation and body-scene alignment without degrading local HMR.

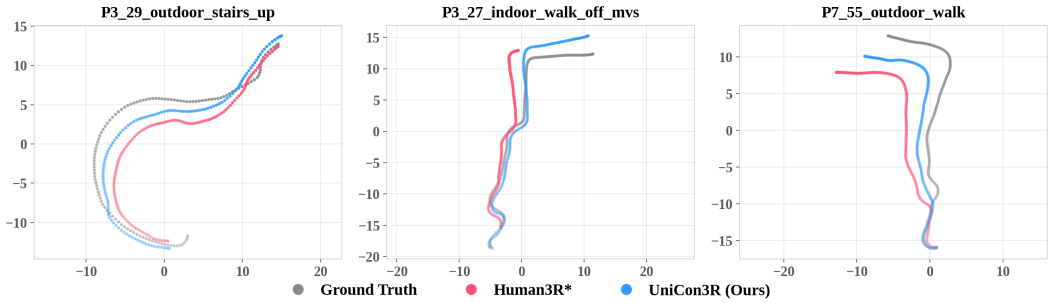

Figure 3: **Comparison of global human trajectory estimation on EMDB-2 [17].** We compare Human3R* and UniCon3R against ground truth after world-coordinate alignment.

Table 2: **Local human mesh reconstruction on SLOPER4D [9] and 3DPW [41].**

Method	SLOPER4D [9]			3DPW [41]	
	PVE (mm) ↓	MPJPE (mm) ↓	Pen. Max (cm) ↓	PVE (mm) ↓	MPJPE (mm) ↓
UniSH [20]	177.0	195.1	725.4	94.2	80.1
Human3R* [7]	<u>155.1</u>	<u>179.1</u>	<u>219.7</u>	**86.6**	<u>72.8</u>
UniCon3R (Ours)	**152.9**	**176.1**	**120.3**	<u>86.8</u>	**72.5**

The qualitative results in Figs. 3 and 4 show the same pattern. On EMDB-2, UniCon3R follows the ground-truth trajectory more closely than Human3R*. On SLOPER4D, it produces body pose and scene alignment that are closer to ground truth, especially in cases with clear scene support.

5.2 Physical Plausibility: Penetration and Floating Meshes

We next evaluate whether the reconstructed body is physically grounded in the recovered scene. UniCon3R reduces **Pen. Max** substantially on SLOPER4D as shown in Table 2. UniCon3R outperforms baselines across all metrics also on RICH as shown in Table 3 (right). Compared with Human3R*, it reduces **Penetrate** by $\sim 75\%$, **Float** by $\sim 83\%$, and **Pen. Max** by $\sim 92\%$. Similar improvements exist over the Parallel Readout baseline. These results show that UniCon3R improves body grounding in world coordinates by reducing scene penetration and unsupported bodies. Note that we do not report Pen. Max on 3DPW as the metric assumes a reliable horizontal ground plane, which is often violated in 3DPW, especially in sequences with non-planar or varying support surfaces.

We complement the findings in Table 3 with qualitative examples on the moving-camera subset of RICH. Fig. 5 compares Human3R (red) and UniCon3R (cyan) against the ground-truth body (gray) on a streaming sequence. Yellow boxes highlight regions where Human3R produces visibly floating or penetrating reconstructions: the body either hovers above the support surface or intersects the scene mesh. UniCon3R places the body closer to the support surface in these same frames, producing more plausible body–scene interaction. The visual pattern matches the quantitative reductions and shows that the contact-guided refinement of UniCon3R resolves the specific failure modes that drive Human3R's grounding errors.

5.3 Contact Evaluation

We evaluate contact prediction on RICH [14] as it provides dense body-scene contact annotations together with video sequences, scene geometry, and human motion. As shown in Table 3 (left), UniCon3R achieves the best F1 score and geodesic error among the compared methods. Compared with *Parallel Readout*, it improves F1 by $\sim 14.5\%$, while reducing geodesic error by $\sim 27.5\%$. Since both models use the same contact supervision, this shows that the gains come from our contact-aware architecture rather than supervision alone. Human3R* does not explicitly predict contact, so we derive contact geometrically at test time by thresholding the distance between each predicted human-mesh vertex and its nearest reconstructed scene point at 0.05 m. We also report DECO [35] and InteractVLM [11] as off-the-shelf dense-contact estimation baselines. Note that both DECO and InteractVLM are single-frame contact estimation methods, whereas UniCon3R estimates contact in the challenging video-streaming setting. While DECO achieves better Precision, UniCon3R obtains the best Recall, F1, and geodesic error. InteractVLM obtains a geodesic error close to UniCon3R but

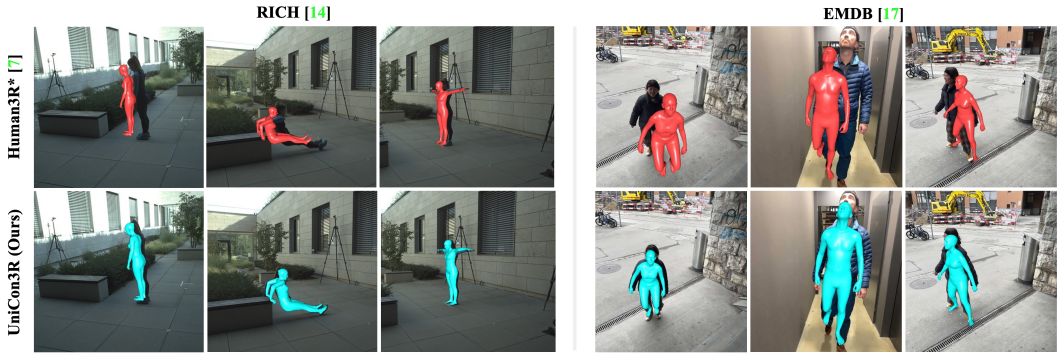

Figure 4: **Qualitative comparison of local human mesh recovery.** We compare Human3R* and UniCon3R against ground truth. UniCon3R produces body pose and scene alignment that are closer to ground truth, particularly in cases with clear scene support.

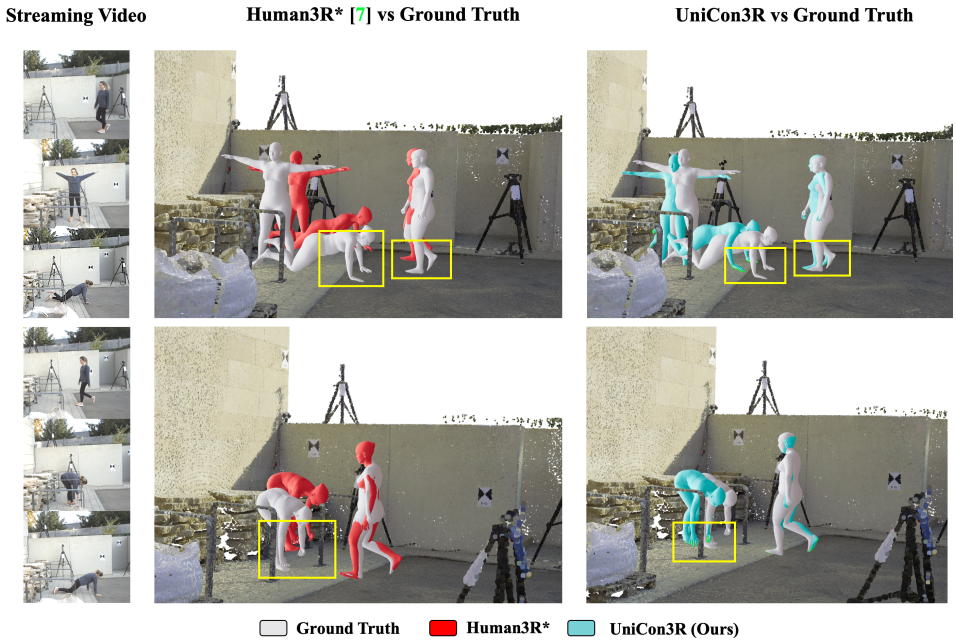

Figure 5: **Qualitative comparison of physical plausibility on RICH.** We compare Human3R and UniCon3R against ground truth on a streaming sequence. Yellow boxes highlight implausible 4D reconstructions of Human3R compared to the more grounded predictions of UniCon3R.

is slower at inference, as discussed in **Sup. Mat.** Qualitatively, Fig. 6 compares contact prediction performance of UniCon3R against DECO on arbitrary internet images. DECO often focuses on feet-ground contact, whereas UniCon3R recovers broader body-scene contact patterns.

In summary, our contact evaluation supports two conclusions. First, contact supervision alone is not sufficient: feeding contact back into the reconstruction pathway improves dense contact prediction over *Parallel Readout*. Second, UniCon3R provides competitive dense contact estimates within a unified reconstruction pipeline, making it 2.0x faster than DECO [35] (See **Sup. Mat.** Section A.3).

5.4 Ablation Studies

We perform a progressive ablation on RICH [14] to evaluate the contribution of each component in UniCon3R. Starting from the *Parallel Readout* baseline, we incrementally add scene context, explicit metric geometry, temporal momentum, and contact-guided latent refinement. We report WA-MPJPE and W-MPJPE for global human motion estimation, and Foot Sliding and Jitter to assess grounding stability over time. Results are summarized in Table 4.

Table 3: **Contact estimation and physical grounding on RICH [14]. Left:** Binary SMPL-vertex contact prediction using precision, recall, F1, and geometric contact error (Geo., cm). **Right:** physical grounding on reconstructed RICH sequences. Following HuMoS [37], we estimate a robust ground height and use a 5 mm tolerance to report Collision Ratio (Coll., %), Penetrate (Pen., cm), Float (cm), and Pen. Max (P.Max, cm).

Method	RICH [14]			
	Prec. ↑	Rec. ↑	F1 ↑	Geo. ↓
DECO [35]	**0.71**	0.76	0.70	17.92
InteractVLM [11]	0.63	0.63	0.60	15.70
Human3R* [7]	0.07	0.13	0.08	76.10
Parallel Readout (Ours)	0.59	0.72	0.62	20.68
UniCon3R (Ours)	0.64	**0.80**	**0.71**	**14.98**

Method	RICH [14]			
	Coll. ↓	Pen. ↓	Float ↓	P.Max ↓
UniSH [20]	7.39	6.01	37.94	67.87
Human3R* [7]	7.40	6.40	33.56	216.97
Parallel Readout (Ours)	7.71	7.56	37.05	254.91
UniCon3R (Ours)	**5.54**	**1.59**	**5.50**	**17.54**

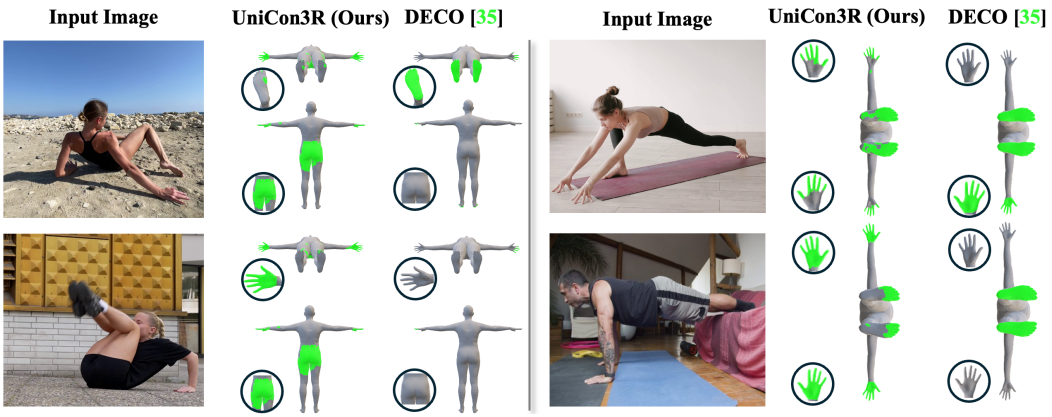

Figure 6: **Qualitative comparison of contact prediction on web images.** UniCon3R captures contact on challenging body regions — such as the lower torso, palms, and fingertips — that are partially occluded or lack direct visual evidence, whereas DECO [35] tends to concentrate predictions on narrower, visually salient support regions such as the feet.

The table also includes Human3R† [7] for reference. Compared with Human3R†, Parallel Readout improves Foot Sliding and Jitter, but slightly worsens WA-MPJPE and W-MPJPE. This shows that contact supervision alone can improve temporal grounding metrics, but is not sufficient to improve global motion estimation. Adding scene context yields the largest single-step improvement over *Parallel Readout*, reducing WA-MPJPE from 97.9 to 89.4 mm and W-MPJPE from 153.5 to 140.6 mm, which highlights the importance of scene-grounded cues for contact reasoning. Adding explicit metric geometry contributes complementary gains, most notably on Foot Sliding (35.0 $\rightarrow$ 33.1 mm), confirming that local 3D support cues further refine grounding. Adding the temporal momentum further improves Foot Sliding and Jitter, indicating better temporal stability. The full model, with contact-guided latent refinement, gives the best result on all four metrics. Relative to *Parallel Readout*, the contact prompt components alone reduce WA-MPJPE from 97.9 to 85.3 mm; adding latent refinement yields a further reduction to 81.5 mm, extending the cumulative gain by 30% and delivering the best result across all metrics. This shows that predicting contact is not enough: the main improvement comes from feeding contact back into the reconstruction pathway. In other words, contact is most useful when it is corrective rather than descriptive.

5.5 UniCon3R with Test-Time Optimization

UniCon3R is designed for online feed-forward inference, and all main comparisons use this setting. Fig. 1(b) additionally reports an optional post-hoc variant, UniCon3R*, which applies a JOSH-style [22] camera refinement while keeping the UniCon3R body prediction fixed. This variant is not part of the default pipeline; it only illustrates the accuracy–runtime trade-off when per-sequence refinement is allowed. UniCon3R achieves 81.5 mm WA-MPJPE$_{100}$ at 2.40 FPS, compared with JOSH (102.6 mm, 0.16 FPS), UniSH (118.1 mm, 2.00 FPS), and Human3R† (97.5 mm, 2.41 FPS); the optional test-time optimized variant UniCon3R * further reduces the error to 72.2 mm.

Table 4: **Ablation study on RICH [14].** We report global motion and physical plausibility metrics while progressively adding each UniCon3R component. WA-MPJPE and W-MPJPE are in mm. Foot sliding is in mm and Jitter in m/s^3, respectively.

Model Variant	Architectural Components				Global Motion		Physical Plausibility	
	Context $\mathbf{U}_{\text{scene}}$	Geo. $\mathbf{G}_t$	Mom. $\mathbf{M}_t$	Latent Ref.	WA-MPJPE ↓	W-MPJPE ↓	Foot Sliding ↓	Jitter ↓
Human3R† [7]	–	–	–	–	97.5	153.2	37.7	279.5
Parallel Readout	✗	✗	✗	✗	97.9	153.5	35.3	262.5
+ Scene Context	✓	✗	✗	✗	89.4	140.6	35.0	260.1
+ Explicit Geometry	✓	✓	✗	✗	85.9	<u>134.5</u>	33.1	244.6
+ Temporal Momentum	✓	✓	✓	✗	<u>85.3</u>	135.1	<u>32.0</u>	<u>231.8</u>
+ Latent Refinement (UniCon3R)	✓	✓	✓	✓	**81.5**	**129.8**	**31.5**	**221.4**

6 Conclusion

We presented **UniCon3R**, a unified feed-forward framework for 4D human-scene reconstruction from monocular video that uses contact inside the reconstruction pathway. By introducing a scene-aware contact prompt and contact-guided latent refinement, UniCon3R improves body grounding while preserving the simplicity and efficiency of unified feed-forward inference. Across benchmarks, the results support the conclusion that in unified feed-forward human-scene reconstruction, predicting contact as an auxiliary output is not sufficient; contact must directly influence the final body estimate. Via the UniCon3R framework, we show that this paradigm results in complementary gains; better grounding, better world-coordinate motion estimation, and local mesh accuracy, all achievable with a single end-to-end model.

Limitations and Future Directions. Several limitations remain. In our work, contact is represented as binary per-vertex labels rather than rich physical quantities such as proximity, contact forces or friction. The quality of recovered contact often depends on the quality of the reconstructed scene geometry, and we observe that scene errors can propagate to contact prediction and body-scene alignment. Future work may explore ways to improve scene reconstruction using contact. Our grounding and plausibility metrics assume rigid support geometry that is approximated by the reconstructed scene; more complex settings such as deformable objects, non-rigid scenes, or fine-grained object interactions are not explicitly modeled. Finally, we observe that augmenting UniCon3R with a post-hoc test-time optimization stage further reduces global trajectory error.

7 Acknowledgements

We would like to acknowledge that computational work involved in this research is partially supported by NUS IT's Research Computing group using grant number NUSREC-HPC-00001.

We also thank Dibyadip Chatterjee, Dr. Sourav Mukhopadhyay and Dr. Aritra Bagchi for their invaluable support and helpful discussions.

Competing Interests. The authors declare no competing financial interests related to this work.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

[2] 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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2024.

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[5] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. volume 36, pages 11454–11468, 2023.

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[7] 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.

[8] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S Lakshmipathy, Agniv Chatterjee, Michael J Black, and Dimitrios Tzionas. Pico: Reconstructing 3d people in contact with objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025.

[9] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 682–692, 2023.

[10] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1323–1333, 2024.

[11] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2d foundational models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22605–22615, 2025.

[12] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] 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*, pages 13274–13285, 2022.

[15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022.

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[17] 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 *International Conference on Computer Vision (ICCV)*, 2023.

[18] 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)*, pages 397–408. IEEE, 2024.

[19] 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*, pages 426–446. Springer, 2024.

[20] 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. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[21] 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*, pages 590–606. Springer, 2022.

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

[23] 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*, pages 851–866. 2023.

[24] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[25] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

[27] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2025.

[28] 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*, pages 10975–10985, 2019.

[29] Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3d humans and environments in tv shows. In *European Conference on Computer Vision*, pages 732–749. Springer, 2022.

[30] Sara Rojas, Matthieu Armando, Bernard Ghamen, Philippe Weinzaepfel, Vincent Leroy, and Gregory Rogez. Hamst3r: Human-aware multi-view stereo 3d reconstruction. *arXiv preprint arXiv:2508.16433*, 2025.

[31] 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.

[32] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022.

[34] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[35] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8001–8013, 2023.

[36] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. URL `https://ipman.is.tue.mpg.de`.

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

[38] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, 2020.

[39] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9720–9729, 2021.

[40] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.

[41] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018.

[42] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[43] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023.

[44] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2024.

[45] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021.

[46] Xiangyu Xu, Lijuan Liu, and Shuicheng Yan. Smpler: Taming transformers for monocular 3d human shape and pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3275–3289, 2023.

[47] Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinghua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Otmar Hilliges. Hsr: holistic 3d human-scene reconstruction from monocular videos. In *European Conference on Computer Vision*, pages 429–448. Springer, 2024.

[48] 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*, pages 1–12, 2025.

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

[50] Fengyuan Yang, Tanuj Sur, Tze Ho Elden Tse, and Angela Yao. Humanba: Human-aware bundle adjustment via global human-camera decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2026.

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

[52] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025.

[53] Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and Laszlo A. Jeni. Onlinehmr: Video-based online world-grounded human mesh recovery. In *CVPR*, 2026.

[54] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1216–1226, 2024.

Appendix

Table of Contents

A Analysis 15
- A.1 Empirical Analysis of the Contact Pathway 15
- A.2 Comprehensive Comparison for Global Human Motion Estimation 16
- A.3 Runtime Efficiency 17
- A.4 Backbone Scaling: Lightweight ViT-S/672 Variant 17

B Implementation Details 18

C Broader Impacts 19

A Analysis

A.1 Empirical Analysis of the Contact Pathway

To complement the downstream reconstruction results, we report a small diagnostic analysis of how contact is used inside UniCon3R. The goal is not to claim that attention alone explains model behavior, but to test whether the contact pathway measurably changes the internal computation that produces the final body estimate. Following common practice in transformer analysis, we pair attention-based observations with explicit controls and a counterfactual intervention on the latent residual branch [1, 6, 24, 25, 38, 40].

Evaluation protocol. We evaluate on the moving-camera subset of RICH [14], which contains 40 sequences. We compare three checkpoints under identical inputs and instrumentation: Human3R†, Parallel Readout, and UniCon3R. Human3R† keeps the original Human3R architecture and uses the same training protocol as UniCon3R, while Parallel Readout adds contact supervision without feeding the contact representation back into the human reconstruction pathway. Relative to these baselines, UniCon3R differs in the two contributions of the main paper: it introduces a *scene-aware contact prompt* $\mathbf{C}_t$, constructed from semantic scene context, explicit metric geometry, and a temporal prior, and it applies *contact-guided latent refinement*, which feeds the refined contact representation back into the HMR pathway before final SMPL-X [28] regression.

Human–contact routing. We first analyze layerwise coupling between the human token and the contact token inside the joint decoder token sequence $[\mathbf{z}, \mathbf{F}_t, \mathbf{H}_t, \mathbf{C}_t]$ processed by the frozen 4D decoders inherited from Human3R [7]/CUT3R [42]. Fig. 7 **(left)** plots the mean coupling in both directions. The goal is not to treat attention weights as explanations by themselves, but to test whether the *scene-aware contact prompt* participates in the same decoder computation as the human prompt. To do so, we compare the human–contact coupling against a matched control derived from ordinary image tokens. Concretely, for each human–contact interaction, we replace the target token by a same-size slice drawn from the image-token set $\mathbf{F}_t$ and measure the corresponding attention mass under the same decoder layer and head. This gives a baseline for how much attention would be assigned to a generic token interaction under otherwise identical conditions. If the human–contact coupling is much larger than this baseline, then the effect is unlikely to be explained by generic attention concentration alone. As observed in Fig. 7 **(left)**, across the moving-camera subset, the strongest $\mathbf{H}_t \to \mathbf{C}_t$ coupling appears in the earliest decoder layer, while the strongest $\mathbf{C}_t \to \mathbf{H}_t$ coupling appears near the final decoder layers. Relative to the matched control, both couplings are substantially larger, indicating a structured interaction pattern rather than generic attention non-uniformity.

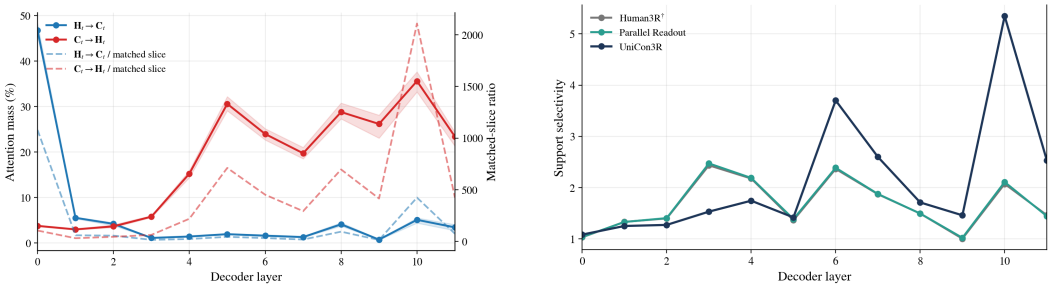

Figure 7: **Empirical Analysis of the Contact Pathway. Left:** layerwise human–contact coupling inside the joint decoder token sequence. Solid curves denote the mean coupling mass (left axis), showing strongest early $\mathbf{H}_t \to \mathbf{C}_t$ coupling and strongest late $\mathbf{C}_t \to \mathbf{H}_t$ coupling. Dashed curves denote the corresponding matched-slice ratios (right axis), where each coupling is normalized by a matched control from a same-size slice of image tokens in $\mathbf{F}_t$. **Right:** support-region selectivity of the human stream. UniCon3R allocates more attention mass to support-relevant image regions than Human3R† and Parallel Readout. Together, these diagnostics support that contact is used inside the reconstruction pathway as an internal corrective signal.

Support-region selectivity. We next test whether the *scene-aware contact prompt* biases the human stream toward image regions that are physically relevant for support. For each frame, we project the ground-truth RICH contact vertices into the image plane to form a binary support-region mask, and define *support selectivity* as the unitless ratio of average attention mass inside the support region to the average mass outside it. Fig. 7 **(right)** plots this ratio across decoder layers for Human3R†, Parallel Readout, and UniCon3R. Averaged over the official dynamic-camera RICH test subset, the three models obtain 1.5, 1.5, and 1.7, respectively, and UniCon3R exceeds both baselines in all camera-sequence entries. The gain is moderate but highly consistent, and indicates that the scene-aware contact prompt reallocates the human-stream attention toward support-relevant image regions.

Counterfactual residual ablation. Finally, we directly ablate the latent residual branch by setting the residual to zero at inference time and re-running only the final SMPL-X [28] regression head. This intervention changes the reconstructed body while leaving the contact prediction unchanged. Averaged over the moving-camera subset, the ablation induces a body-root translation change of $\|\Delta\mathbf{\Gamma}\|_2 = 0.3\,\text{m}$ (here $\mathbf{\Gamma}$ denotes the world-frame SMPL-X root translation, i.e., $\mathbf{P}_t$ in the main paper notation), a mean per-joint angular shift of $2.8°$, and a maximum per-joint shift of $10.2°$, while the contact-logit difference remains 0.0. This supports distinct roles for the two branches of the contact pathway: the contact head predicts dense contact, whereas the latent refinement branch uses the contact-conditioned representation to modify the final body configuration.

Taken together, these diagnostics support the same conclusion as the main experiments. In UniCon3R, contact is not used as a detached auxiliary output. Instead, it is routed through a dedicated contact representation and then fed back into the human stream through contact-guided latent refinement, where it acts as an internal corrective signal for the final body estimate.

A.2 Comprehensive Comparison for Global Human Motion Estimation

Table 5 provides the comprehensive comparison for global human motion estimation on RICH [14] and EMDB-2 [17]. In the main paper, we focus on the *Unified* setting, where methods jointly recover humans, camera motion, and dense scene geometry from monocular video without per-sequence optimization. This is the setting most directly matched to UniCon3R. For completeness, we also include methods from two related settings.

We group methods by inference setting. **Offline** methods use per-sequence optimization or test-time optimization, and can therefore adapt predictions to the full test sequence. **World-HMR** methods estimate global human motion in world coordinates, but do not jointly reconstruct dense scene geometry in the same unified feed-forward model. **Unified** methods jointly predict humans, camera motion, and scene geometry in a feed-forward reconstruction pipeline. Since these categories differ in

Table 5: **Comprehensive Comparison for Global human motion estimation on RICH [14] and EMDB-2 [17]. Scene** denotes 3D scene reconstruction, and **Contact** denotes dense human-scene contact prediction or use. Human3R*: released checkpoint in our pipeline; Human3R†: same architecture with our training protocol. **Bold** and underline indicate the best and second-best results within the *Unified* category, where UniCon3R operates. WA-MPJPE, W-MPJPE, and PA-MPJPE are reported in mm; RTE is reported in %.

Category	Method	Scene	Contact	RICH [14]				EMDB-2 [17]			
				WA-MPJPE ↓	W-MPJPE ↓	RTE (%) ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE (%) ↓	PA-MPJPE ↓
Offline	SLAHMR [51]	✗	✗	132.2	237.1	6.4	–	326.9	776.1	10.2	61.5
	TRAM [44]	✗	✗	127.8	238.0	6.0	–	76.4	222.4	1.4	38.1
	JOSH [22]	✓	✓	102.6	184.3	3.4	–	68.9	174.7	1.3	–
World-HMR	TRACE [34]	✗	✗	238.1	925.4	101.4	–	529.0	1702.3	17.7	58.0
	WHAM [32]	✗	✗	108.4	190.1	4.5	–	135.6	354.8	6.0	38.2
	GVHMR [31]	✗	✗	78.8	126.3	2.4	–	111.0	276.5	2.0	–
	OnlineHMR [53]	✓	✗	–	–	–	–	93.5	310.4	2.2	40.1
Unified	JOSH3R [22]	✓	✗	190.4	334.9	6.3	–	220.0	661.7	13.1	–
	UniSH [20]	✓	✗	118.1	183.2	4.8	50.0	118.5	**270.1**	5.8	44.4
	Human3R* [7]	✓	✗	109.9	184.0	3.3	48.6	<u>118.2</u>	286.1	<u>2.4</u>	<u>38.7</u>
	Human3R† [7]	✓	✗	<u>97.5</u>	<u>153.2</u>	<u>2.3</u>	<u>41.4</u>	–	–	–	–
	Parallel Readout (Ours)	✓	✓	97.9	153.5	<u>2.3</u>	41.5	160.7	388.3	3.6	50.9
	UniCon3R (Ours)	✓	✓	**81.5**	**129.8**	**1.7**	**37.6**	**113.7**	<u>285.6</u>	**2.3**	**38.2**

task scope and inference procedure, comparisons across categories should be interpreted as contextual rather than controlled.

Within the **Unified** category, UniCon3R achieves the best results on RICH across WA-MPJPE, W-MPJPE, RTE, and PA-MPJPE. On EMDB-2, it improves over all Human3R-derived baselines and achieves the best WA-MPJPE, RTE, and PA-MPJPE within the same category. Compared with Human3R†, the gains are not explained by the training protocol alone; compared with Parallel Readout, they show that contact is most useful when it is fed back into the reconstruction pathway rather than predicted as an auxiliary output.

The broader comparison also shows the remaining gap between unified feed-forward reconstruction and methods that use offline or post-hoc refinement. Some Offline and World-HMR methods obtain competitive global motion accuracy, but they do not address the same unified online human-scene reconstruction problem as UniCon3R. Our focus is therefore on improving global motion and physical grounding within a single feed-forward human-scene reconstruction pipeline.

A.3 Runtime Efficiency

In all our experiments, Human3R* denotes our evaluation of the released Human3R checkpoint (ViT-L/896 variant) under the same preprocessing, inference, and evaluation pipeline used for UniCon3R. For end-to-end reconstruction, Human3R* [7] runs at 2.41 FPS on a single NVIDIA A5000 GPU, while UniCon3R runs at 2.40 FPS under the same preprocessing and benchmarking protocol. This shows that the contact branch adds negligible overhead to the full reconstruction pipeline.

For contact prediction, UniCon3R runs at 0.5 s/frame, compared with 1.1 s/frame for DECO [35] and 3.5 s/frame for InteractVLM [11]. Thus, UniCon3R is about $2.0\times$ faster than DECO and about $6.1\times$ faster than InteractVLM on this task. Unlike these methods, contact in UniCon3R is integrated into the reconstruction pipeline rather than applied as a separate post-processing stage.

A.4 Backbone Scaling: Lightweight ViT-S/672 Variant

The main paper instantiates UniCon3R on top of the largest Human3R [7] checkpoint (ViT-L/896 with Multi-HMR [2] ViT-L), which delivers the strongest accuracy at the cost of inference speed. Because Human3R itself is released as a family of backbones (ViT-S/672, ViT-B/672, ViT-L/672, ViT-L/896) that trade accuracy for speed, a natural question is whether the contact-aware design of UniCon3R continues to help when the underlying backbone is much smaller. We address this by also instantiating UniCon3R on the lightweight ViT-S/672 checkpoint, in which both the input resolution (672×672) and the Multi-HMR encoder size (ViT-S) are reduced. The two contributions of UniCon3R—the *scene-aware contact prompt* and the *contact-guided latent refinement*—are unchanged. The training protocol (RICH [14] + BEDLAM [3] fine-tuning, see Section B) is also identical to the main paper.

Table 6: **Global human motion estimation across backbone scales on RICH [14] and EMDB-2 [17].** We compare the lightweight ViT-S/672 variant and the ViT-L/896 variant used in the main paper. **Scene** denotes 3D scene reconstruction; **Contact** denotes dense human-scene contact prediction or use. Human3R*: released checkpoint at the indicated backbone, evaluated in our pipeline; Human3R†: same architecture and backbone with our training protocol. WA-MPJPE, W-MPJPE, and PA-MPJPE are in mm; RTE in %. **Bold**/underline denote best/second-best *within each backbone group*.

Method	Backbone	Scene	Contact	RICH [14]				EMDB-2 [17]			
				WA-MPJPE ↓	W-MPJPE ↓	RTE (%) ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE (%) ↓	PA-MPJPE ↓
Human3R* [7]	ViT-S/672	✓	✗	131.5	207.3	3.3	71.8	137.8	**333.1**	**2.4**	**52.9**
Human3R† [7]	ViT-S/672	✓	✗	113.9	178.3	2.6	58.3	–	–	–	–
UniCon3R (Ours)	ViT-S/672	✓	✓	**97.2**	**151.1**	**2.0**	**51.7**	**130.7**	331.6	2.5	**52.9**
Human3R* [7]	ViT-L/896	✓	✗	109.9	184.0	3.3	48.6	118.2	286.1	2.4	38.7
Human3R† [7]	ViT-L/896	✓	✗	97.5	153.2	2.3	41.4	–	–	–	–
UniCon3R (Ours)	ViT-L/896	✓	✓	**81.5**	**129.8**	**1.7**	**37.6**	**113.7**	**285.6**	**2.3**	**38.2**

Under the lightweight configuration, UniCon3R runs at 5.64 FPS on a single NVIDIA A5000 GPU (mean over the RICH moving-camera test subset), compared with 5.85 FPS for the corresponding Human3R [7] ViT-S/672 backbone evaluated under the same protocol. The ~ 0.21 FPS overhead is attributable to the additional contact branch and reinforces the negligible-overhead behavior reported at the ViT-L/896 scale in Section A.3. Compared with the ViT-L/896 UniCon3R from the main paper (2.40 FPS), the lightweight variant is therefore approximately $2.4\times$ faster, making it the appropriate choice when reduced compute budget is the primary requirement.

Merged comparison across backbones. Table 6 compares the two backbone variants of UniCon3R side by side against backbone-matched Human3R baselines: Human3R*, the released checkpoint at each scale, evaluated in our pipeline; and Human3R†, the same architecture and backbone fine-tuned with our training protocol so that supervision is matched. Because the two backbones differ substantially in capacity, we apply the bold/underline convention *within each backbone group* rather than across groups.

Global human motion estimation. On global motion (Table 6), the contact pathway delivers consistent improvements at *both* backbone scales. At ViT-S/672, UniCon3R reduces WA-MPJPE on RICH from 113.9 mm (Human3R†) to 97.2 mm (-14.6%), and W-MPJPE from 178.3 mm to 151.1 mm (-15.3%). At ViT-L/896, the corresponding reductions are $97.5 \rightarrow 81.5$ mm (-16.4%) on WA-MPJPE and $153.2 \rightarrow 129.8$ mm (-15.3%) on W-MPJPE. The relative improvement from contact-aware refinement is therefore stable across the $\sim 2.4\times$ speed difference between the two variants.

When to use which variant. Together, Table 6 supports a clean recommendation. The lightweight ViT-S/672 variant is the appropriate choice when compute budget is the primary constraint: it preserves the world-coordinate gains of contact-guided refinement at 5.64 FPS, roughly $2.4\times$ the throughput of the ViT-L/896 variant. For applications where fine-grained body-mesh detail matters most—such as expressive avatar reconstruction or close-range human-scene interaction analysis—the ViT-L/896 variant from the main paper remains the better choice. Importantly, the contact-aware design of UniCon3R transfers cleanly across backbone scales without any architecture change, indicating that the proposed contact pathway is complementary to, rather than competing with, the backbone-scaling axis explored by Human3R [7].

B Implementation Details

Backbone and initialization. We build UniCon3R on top of the released Human3R [7] checkpoint, which adapts CUT3R [42] through visual prompt tuning. As in Human3R, we keep the pretrained CUT3R backbone frozen during fine-tuning, and we also keep the Multi-HMR [2] encoder frozen. The frozen scene backbone uses a ViT-Large image encoder with token dimension 1024, followed by two interleaved 4D decoders inherited from CUT3R with embedding dimension c=768, depth 12, and 12 attention heads. The persistent state $\mathbf{S}_t$ also has dimension 768.

Trainable modules. We optimize only the human- and contact-related modules introduced by UniCon3R. These include the scene-context cross-attention layers, the prompt-construction MLPs used to form the *scene-aware contact prompt*, the contact head, the residual head $\mathrm{Head}_{\mathrm{residual}}$ used for *contact-guided latent refinement*, the SMPL-X prediction head, and the segmentation head inherited from Human3R. Unless noted otherwise, all MLPs use two linear layers with GELU activations and hidden dimension 768. The SMPL-X parameter decoders inside $\mathrm{Head}_{\mathrm{human}}$ follow Human3R and use separate two-layer MLPs with hidden dimension 4×1792 to regress body pose, shape, expression, and root translation from $\bar{\mathbf{H}}_t$.

Local geometry token. The explicit metric geometry cue in Sec. 4.1 is constructed from the previous-frame world-frame pointmap $\mathbf{X}_{t-1}$ using the same notation as in the main paper. For each detected human anchor $\mathbf{u}_t^n$, we apply RoIAlign [13] to a square window centered at $\mathbf{u}_t^n$ with half-size 24 pixels and output size 7×7. The pooled 3-channel coordinates are then averaged to obtain the local descriptor $\phi_{\mathrm{geo}}(\mathbf{X}_{t-1}, \mathbf{u}_t^n) \in \mathbb{R}^3$, which is mapped into the decoder space to form the geometry token used in $\mathbf{G}_t$.

Training data. We fine-tune on a balanced mixture of RICH [14] and BEDLAM [3]. Following CUT3R and Human3R, we exclude BEDLAM sequences whose environment is represented as a panoramic HDRI image, yielding 2700 training sequences. RICH provides multi-camera video with scanned scene geometry, SMPL-X meshes, and dense body-scene contact annotations; we use the official split of RICH sequences leaving the moving camera subset for evaluation.

Optimization. Training uses AdamW with β_1=0.9, β_2=0.95, and weight decay 0.05. We use a linear warmup over the first 5 epochs from 10^{-6} to 10^{-5}, followed by cosine decay back to 10^{-6}, and train for 100 epochs. The per-GPU batch size is 4 with gradient accumulation of 2, giving an effective batch size of 64 on 8 NVIDIA A5000 GPUs. We use mixed precision and gradient checkpointing.

Evaluation Metrics For evaluating global human motion, we split sequences into segments of 100 frames and align the predicted motion with the ground truth using either the first two frames or the entire segment and compute the mean-per-joint error in millimeters to obtain W-MPJPE or WA-MPJPE for evaluation. We also evaluate the relative Root Translation Error (RTE%) of the whole trajectory.

Training objective. The total training objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{4D}} + \mathcal{L}_{\mathrm{SMPLX}} + \lambda_c \mathcal{L}_{\mathrm{contact}}.$$

For the inherited reconstruction pathway, we use the CUT3R confidence-aware $L_{2,1}$ pointmap loss with α=0.2 together with the per-frame RGB MSE loss, and supervise the human output with the Human3R L_1 SMPL-X loss. For contact, we use focal BCE with γ=2.0 and α=0.25, a positive-class weight of 10, and a part-level loss weighted by λ_p=1.5, with overall contact weight λ_c=1.0. Since BEDLAM does not provide body-scene contact labels, contact supervision is applied only on RICH samples, while BEDLAM samples optimize only the inherited reconstruction and HMR objectives.

Contact supervision space. Following BSTRO [14], we supervise contact in SMPL [23] space rather than SMPL-X space. The human regressor remains in SMPL-X space for compatibility with Human3R, while the contact head predicts dense contact on the 6890 SMPL vertices, where most body-scene support contact occurs.

C Broader Impacts

UniCon3R improves the reconstruction of human motion and human-scene interaction from monocular video by using contact to refine body pose and scene alignment. More physically plausible reconstructions may benefit applications such as robotics, human motion analysis, ergonomics, and immersive AR/VR, where understanding how people interact with their surroundings is important.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three claims—(i) treating contact as an internal corrective signal, (ii) the *scene-aware contact prompt* and *contact-guided latent refinement* components, and (iii) gains in physical plausibility, global motion, and local mesh recovery while preserving feed-forward inference—and each is supported by the experiments in Sections 5.1 to 5.4 and the runtime results in Section A.3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses binary per-vertex contact labels, dependence on reconstructed scene geometry, and limitations involving deformable objects, non-rigid scenes, and fine-grained interactions. The default method does not require test-time optimization; Section 5.5 separately evaluates an optional post-hoc refinement variant.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical framework for 4D human-scene reconstruction and does not include formal theorems, lemmas, or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies the contact-prompt design, joint decoder integration, residual refinement, and training objectives. Section B provides the backbone, trainable modules, optimizer, learning-rate schedule, batch sizes, data sources, and loss hyperparameters. The experimental section describes the reported metrics, with additional global-motion evaluation details in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets used in the experiments are available from their original providers under the applicable access conditions and licenses. The source code and trained model checkpoints are not yet publicly released with this manuscript. Nevertheless, Section 4 and Section B document the architecture and training procedure to support independent reproduction. We intend to release the implementation and trained checkpoints.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section B reports the AdamW optimizer, learning-rate schedule, batch sizes, loss weights, RICH/BEDLAM training setup, and held-out moving-camera RICH evaluation subset. It also specifies the 100-frame segmentation and alignment protocol for global human motion evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Following standard practice in monocular 4D human-scene reconstruction [7, 20, 22], we report point estimates on the official test splits of RICH, EMDB-2, 3DPW, and SLOPER4D rather than error bars over re-training seeds, because each fine-tuning run on 8 A5000 GPUs is computationally expensive and the test sets are fixed and deterministic. Trends are nevertheless consistent across two backbone scales (Table 6) and across four datasets, and the controlled comparison against Human3R† and Parallel Readout isolates the effect of our contributions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section B reports training on eight NVIDIA A5000 GPUs for 100 epochs with mixed precision and gradient checkpointing, while Sections A.3 and A.4 report single-GPU inference throughput. However, the manuscript does not yet report per-run training duration, GPU-hours, or an estimate of total experimental and development compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics. This work uses previously released datasets and does not involve new human-subjects recruitment or data collection by the authors. The broader-impacts discussion addresses potential privacy and surveillance risks associated with downstream use of monocular human-scene reconstruction.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix C discusses the potential benefits of contact-aware human-scene reconstruction for robotics, human motion analysis, ergonomics, and immersive AR/VR, where accurate understanding of human-scene interaction can support downstream applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The proposed model predicts 3D human-scene reconstruction and body-scene contact; it does not generate synthetic media or introduce a new scraped dataset. We do not identify a need for the high-risk asset-release safeguards described in this question. Privacy and surveillance risks associated with possible downstream use are addressed in Appendix C.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper credits the original creators of the datasets and pretrained models used, including RICH, BEDLAM, EMDB, 3DPW, SLOPER4D, CUT3R, Human3R, Multi-HMR, SMPL, and SMPL-X. However, the manuscript does not yet explicitly document exact licenses for each asset, applicable version, and access terms. These details should be verified against the original releases and added before claiming full compliance.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This manuscript describes the new UniCon3R architecture and documents its components and training settings in Sections B and 4, but it does not currently release a new dataset, model checkpoint, or code package as an accompanying public asset. Code, checkpoints, and related documentation are intended for a subsequent release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not conduct any crowdsourcing or new human-subjects data collection; it only uses existing publicly released datasets that were captured and annotated by their original authors.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human-subjects research was conducted by the authors; all human data comes from existing public datasets whose creators handled the corresponding IRB/consent procedures.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the proposed method—no LLM is used for contact prediction, scene reconstruction, human mesh recovery, or any other core element of UniCon3R.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.