METAV3RSE: 4D HUMAN-OBJECT-SCENE RECONSTRUCTION FROM MONOCULAR VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **METAV3RSE**, a unified feed-forward framework that reconstructs the human, the manipulated object, and the surrounding scene together in one online pass from monocular video. Existing feed-forward methods address either human-object or human-scene reconstruction, not the coupled setting, so object recovery cannot use the human's scene grounding and the scene-grounded human cannot use the constraints supplied by the object. Methods that do reconstruct all three rely on per-sequence optimization. The difficulty is that the manipulated object is the least visible part of the interaction. The reconstructed human constrains where the object lies but not its precise rigid pose, while visible object evidence is precise but incomplete. We therefore extend the recurrent human-scene state with object and human-object contact representations that produce an interaction-conditioned pose hypothesis and feed interaction back to the human, then correct that hypothesis using visual evidence gathered at canonical points of the known object template. We show that **METAV3RSE** reduces combined human-object reconstruction error by 24% over the strongest feed-forward baseline on BEHAVE, and joint object reasoning also improves the human, reducing WA-MPJPE by over 23% on the unseen EMDB-2 benchmark.

1 INTRODUCTION

Everyday human activity couples object manipulation with movement and contact with the surrounding scene. We refer to this setting as **human-object-scene interaction (HOSI)**, bringing together human-object interaction (HOI) and human-scene interaction (HSI). Existing methods typically address either human-scene or human-object reconstruction, leaving complementary constraints disconnected. We posit that joint reconstruction allows the scene-grounded body to guide object placement under occlusion, while object interaction refines the human in return.

Joint HOSI reconstruction is challenging because the human, object, and scene provide different but complementary constraints. First, the manipulated object is often the least visible part of the interaction. Under heavy occlusion, the visible object may be consistent with several rigid poses, making appearance alone insufficient, while the reconstructed human, its motion, and human-object interaction provide strong cues for where the object should lie (Xie et al., 2023; Cseke et al., 2025; Zhang et al., 2026). These cues are nevertheless incomplete: errors in body pose or contact can leave residual misalignment that visible object evidence is needed to correct.

Secondly, the human, object, and scene must remain mutually consistent even though each is recovered from incomplete monocular evidence. Human-scene methods (Sur et al., 2026; Li et al., 2026; Chen et al., 2026; Liu et al., 2026) ground the body in the surrounding scene but do not explicitly reconstruct the manipulated object, while feed-forward HOI methods (Nam et al., 2024; Wang et al., 2025b; Zhang et al., 2026) reconstruct the human-object pair without the same scene-grounded recurrent representation. Methods that couple interaction and object evidence through optimization (Cseke et al., 2025; Xie et al., 2023; 2026) require test-time fitting and are less suited to direct online inference. The challenge is therefore not merely to reconstruct three components together, but to make their interaction actively shape the reconstruction within a single feed-forward model.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this, we introduce **METAV3RSE**, a recurrent feed-forward framework[1] for unified 4D HOSI reconstruction. Our first component, *interaction-aware object initialization*, extends the contact-aware human-scene reconstruction of UniCon3R (Sur et al., 2026) with object and human-object contact representations in the same recurrent context. Importantly, we also extend UniCon3R's contact-guided latent refinement principle. UniCon3R uses human-scene contact not only as an output, but as a corrective signal that updates the human representation before body regression. We introduce the missing object-to-human counterpart: local human-object interaction produces a residual update to the human representation while simultaneously conditioning the object representation and human-object contact. The coupling is therefore bidirectional: the scene-grounded human provides a strong reference for object placement, while object interaction feeds back to refine the human itself. From this shared representation, we predict human-object contact and an interaction-conditioned object-pose hypothesis relative to the reconstructed body.

The interaction-conditioned hypothesis provides a useful object placement when direct appearance is incomplete. But interaction alone does not determine the final pose. A second component, the *mesh-indexed object pose readout*, uses the hypothesis to determine where object-specific visual evidence should be gathered. Canonical points sampled from the known object template are projected into the image under the initial pose, so that each template point queries visual evidence at its expected location. Combining this evidence with the interaction-conditioned object representation allows the network to predict corrected camera-space locations for the template points. Because the object is rigid, these predictions are consolidated into a single rotation and translation through a closed-form Kabsch readout (Kabsch, 1978; Huang et al., 2022b). In this way, interaction provides the initial placement, object appearance corrects the remaining pose ambiguity, and the final object remains rigid. All components operate within a single forward pass, extending recurrent human-scene reconstruction to the manipulated object without per-frame or per-sequence pose optimization.

In summary, our contributions are:

- We formulate **4D human-object-scene interaction (HOSI) reconstruction** as a unified online perception problem, jointly recovering the human, manipulated object, surrounding scene, camera motion, and their contacts in a shared metric world frame, and introduce **METAV3RSE** as a feed-forward solution without per-sequence optimization.
- We extend contact-guided latent refinement beyond human-scene interaction to the manipulated object, closing the interaction loop: the scene-grounded human guides object placement, while human-object interaction feeds back to refine the human itself.
- We introduce an interaction-aware object initialization and mesh-indexed pose readout, where the interaction-conditioned pose determines where canonical object points gather visual evidence to consolidate into the final rigid pose through a closed-form readout.

2 RELATED WORK

Human Mesh, Motion, and Human-Scene Reconstruction. Monocular human reconstruction spans per-image pose and shape estimation (Kanazawa et al., 2018; Kocabas et al., 2021; Li et al., 2022; Goel et al., 2023), temporally consistent video methods (Kocabas et al., 2020; Choi et al., 2021), and world-grounded motion recovery under moving cameras (Yuan et al., 2022; Shin et al., 2024; Shen et al., 2024; Yang et al., 2025; 2026). Recent feed-forward human-scene methods (Liu et al., 2026; Chen et al., 2026; Li et al., 2026) jointly recover the human, dense scene geometry, and camera motion in a common coordinate frame. Within this line of work, recurrent models maintain scene context across frames. UniCon3R (Sur et al., 2026) further uses predicted human-scene contact to refine the body, demonstrating that interaction cues can improve reconstruction rather than serve only as outputs. These methods establish scene-grounded human reconstruction but do not explicitly recover manipulated objects. METAV3RSE builds upon these works to introduce object-pose estimation and human-object interaction feedback within one common framework.

Human-Object Reconstruction. Human-object reconstruction uses human context to constrain the ambiguous object poses under occlusion. Optimization-based methods fit object templates using

[1]Following CUT3R (Wang et al., 2025a), we use *feed-forward* to denote direct inference without per-sequence optimization, although the model maintains a recurrent state.

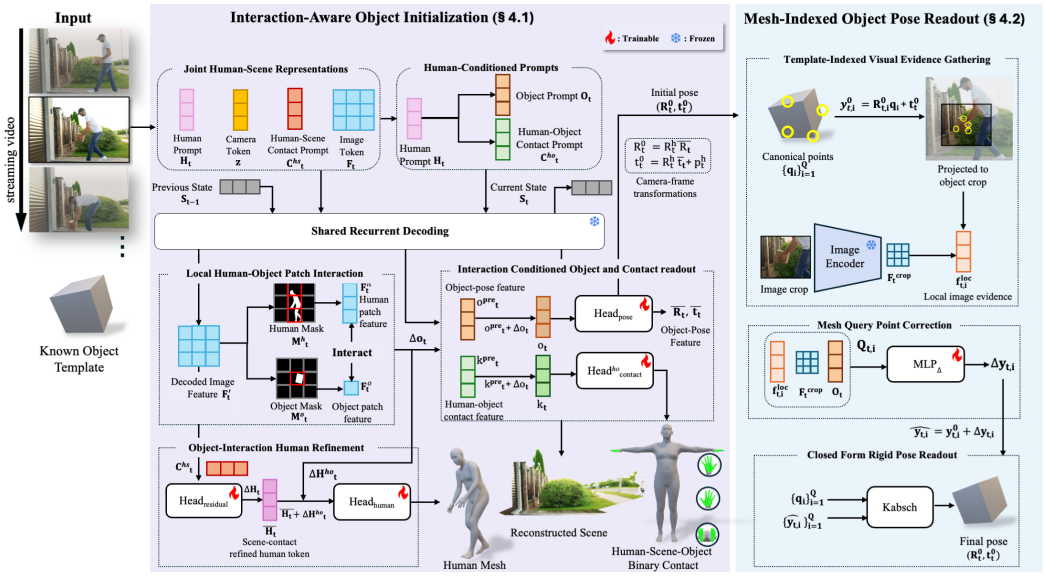


Figure 1: **Overview of METAV3RSE. Left:** the human representation initializes object and human-object contact prompts, decoded alongside the inherited tokens and recurrent state. Masked human and object patches exchange information, yielding a human residual $\Delta\mathbf{H}_t^{ho}$ that refines the body before SMPL-X regression and an object residual $\Delta\mathbf{o}_t$ that drives human-object contact and an initial object pose (Section 4.1). **Right:** this pose projects canonical template points into the object crop, where mesh queries gather image evidence and predict 3D corrections. Closed-form Kabsch alignment (Kabsch, 1978) consolidates them into the final rigid object pose (Section 4.2).

silhouettes, learned distance fields, and contact relations (Zhang et al., 2020; Xie et al., 2022), while feed-forward methods learn human-object dependencies within the network (Nam et al., 2024; Wang et al., 2025b; Chatterjee & Pavlakos, 2026). CONTHO, for example, uses estimated contact to guide feature exchange between the human and object. Such interaction cues constrain object placement, but incomplete or uncertain estimates leave ambiguities that visible object evidence can help resolve.

A complementary strategy associates image evidence with known object geometry. CONTHO samples features at projected vertices of the object estimate, while HOI-TG (Wang et al., 2025b) uses projected template geometry to gather features and apply a rigid fitting. METAV3RSE builds on these strategies, using an interaction-conditioned pose to guide template-indexed visual corrections. In our formulation, human-object interaction is bidirectional: human and object features exchange information, and the resulting human-side residual feeds back into the scene-grounded human representation before body regression. The object is thus recovered together with an interaction-refined human in the shared world frame.

4D Human-Object Reconstruction in Scenes. Video methods extend the human cue over time. VisTracker (Xie et al., 2023) and THO (Zhang et al., 2026) infer the object in occluded frames from human motion and neighboring frames, and CARI4D (Xie et al., 2026) enforces predicted contacts during joint optimization. Temporal context strengthens the interaction cue, but these methods reconstruct the human-object pair without the surrounding scene; VisTracker and THO operate in camera coordinates, with THO assuming a static camera. HOSNeRF (Liu et al., 2023) and RHINO (Xue et al., 2026) do reconstruct humans, manipulated objects, and scenes together, but build and refine a scene-specific representation over each sequence. METAV3RSE targets the online, feed-forward setting: it inserts a known rigid object into an evolving human-scene reconstruction, anchoring it to a body already placed in the world, without per-sequence refinement.

3 PRELIMINARIES

Problem setting. Given a monocular RGB video $\{I_t\}_{t=1}^{T}$ of a person interacting with a single rigid object and a known metric object template $\mathbf{O}$, we recover the human, manipulated object,

scene geometry, and camera motion in a common metric world frame, together with human-scene contacts and human-object contacts over time.

At frame t, the human is represented by SMPL-X (Pavlakos et al., 2019) parameters $\mathbf{Y}_t$, comprising body pose, shape, global orientation, and camera-frame root translation. The camera pose $\mathbf{T}_t \in \mathrm{SE}(3)$ maps camera coordinates to the world frame. The object pose is represented in the camera frame by a rotation $R_t^o \in \mathrm{SO}(3)$ and translation $\mathbf{t}_t^o \in \mathbb{R}^3$. For a canonical template point $\mathbf{o} \in \mathbf{O}$,

$$\mathbf{o}_t^c = R_t^o \mathbf{o} + \mathbf{t}_t^o, \qquad \begin{bmatrix} \mathbf{o}_t^w \\ 1 \end{bmatrix} = \mathbf{T}_t \begin{bmatrix} \mathbf{o}_t^c \\ 1 \end{bmatrix}. \tag{1}$$

Since $\mathbf{O}$ is a known metric rigid template, object reconstruction reduces to estimating its rotation and translation while preserving its shape and scale.

We follow UniCon3R (Sur et al., 2026) and represent contact types on the 6,890-vertex SMPL model: $\hat{\mathbf{c}}_t^{hs}, \hat{\mathbf{c}}_t^{ho} \in [0,1]^{6890}$. The human-object output therefore identifies body vertices in contact with the manipulated object; we do not predict a separate dense contact map on the object template.

Recurrent human-scene reconstruction. We build on UniCon3R (Sur et al., 2026), which adds human-scene contact reasoning to the reconstruction of Human3R (Chen et al., 2026). At frame t, the inherited decoder processes image tokens $\mathbf{F}_t$, human prompt $\mathbf{H}_t$, human-scene contact prompt $\mathbf{C}_t^{hs}$, and camera prompt $\mathbf{z}$ with persistent scene state $\mathbf{S}_{t-1}$:

$$[\mathbf{F}'_t, \mathbf{z}'_t, \mathbf{H}'_t, \mathbf{C}'^{hs}_t], \mathbf{S}_t = \text{Decoders}\left([\mathbf{F}_t, \mathbf{z}, \mathbf{H}_t, \mathbf{C}_t^{hs}], \mathbf{S}_{t-1}\right). \tag{2}$$

The outputs comprise decoded image features $\mathbf{F}'_t$, camera features $\mathbf{z}'_t$ for motion recovery, scene-grounded human features $\mathbf{H}'_t$, and human-scene contact features $\mathbf{C}'^{hs}_t$. Throughout the paper, blue denotes modules inherited from the frozen backbone, and red denotes task-specific prediction and interaction modules introduced or adapted beyond it.

Like UniCon3R, we augment the decoded human and contact features with human-prior features $\mathbf{F}_{\mathrm{HMR},t}$ from the predicted human anchor of a frozen Multi-HMR (Baradel et al., 2024) encoder:

$$\widetilde{\mathbf{H}}_t = \mathbf{H}'_t \oplus \mathbf{F}_{\mathrm{HMR},t}, \qquad \widetilde{\mathbf{C}}_t^{hs} = \mathbf{C}'^{hs}_t \oplus \mathbf{F}_{\mathrm{HMR},t}, \tag{3}$$

where $\oplus$ denotes channel-wise concatenation. The human-scene contact representation predicts dense contact and is also fed back into the human pathway before final SMPL-X regression:

$$\bar{\mathbf{H}}_t = \widetilde{\mathbf{H}}_t + \text{Head}_{\text{residual}}(\widetilde{\mathbf{C}}_t^{hs}), \quad \hat{\mathbf{c}}_t^{hs} = \sigma\left(\text{Head}_{\text{contact}}^{hs}(\widetilde{\mathbf{C}}_t^{hs})\right). \tag{4}$$

Note this inherited reconstruction does not explicitly represent manipulated objects or human-object interaction. Section 4.1 will introduce these representations.

4 METHODOLOGY

As shown in Fig. 1, METAV3RSE extends the contact-aware human-scene reconstruction of Section 3 to reconstruct objects through two components. First, we apply an interaction-aware initialization to predict human-object contact and a body-relative object pose (Section 4.1). We then add a mesh-indexed readout that uses this pose to gather template-specific image evidence and recover a refined rigid pose (Section 4.2). Section 4.3 describes the training procedure. Both components run sequentially during online inference, without per-frame or per-sequence object-pose optimization.

4.1 INTERACTION-AWARE OBJECT INITIALIZATION

We introduce object and human-object contact prompts initialized from the human representation $\mathbf{H}_t$: $\mathbf{O}_t = \mathbf{H}_t + \mathbf{e}_o + \mathrm{CA}(\mathbf{H}_t + \mathbf{e}_o, \mathbf{F}_t)$ and $\mathbf{C}_t^{ho} = \mathbf{H}_t + \mathbf{e}_{ho}$. Learned embeddings $\mathbf{e}_o$ and $\mathbf{e}_{ho}$ distinguish the prompts, while cross-attention CA retrieves image evidence for the object prompt.

The inherited recurrent decoder processes these prompts jointly with human-scene representations:

$$[\mathbf{F}'_t, \mathbf{z}'_t, \mathbf{H}'_t, \mathbf{C}'^{hs}_t, \mathbf{C}'^{ho}_t, \mathbf{O}'_t], \mathbf{S}_t = \text{Decoders}\Big([\mathbf{F}_t, \mathbf{z}, \mathbf{H}_t, \mathbf{C}_t^{hs}, \mathbf{C}_t^{ho}, \mathbf{O}_t], \mathbf{S}_{t-1}\Big). \tag{5}$$

The shared recurrent context provides object-level interaction information, but interaction does not yet enforce consistency between the local human and object regions. We therefore use the predicted human mask $\mathbf{M}_t^h$, obtained from the segmentation head of the human branch, and supplied object mask $\mathbf{M}_t^o$ to select human and object patch features $\mathbf{F}_t^h$ and $\mathbf{F}_t^o$ from $\mathbf{F}_t'$. A bidirectional interaction module learns which human and object patches are related and produces residuals $(\Delta\mathbf{H}_t^{ho}, \Delta\mathbf{o}_t) = \mathrm{Interact}(\mathbf{F}_t^h, \mathbf{F}_t^o)$. The human-object interaction module, Interact, is detailed in the **Sup. Mat.**

We add the human residual to UniCon3R's scene-contact-refined representation before body regression, allowing object interaction to also refine the human:

$$\mathbf{H}_t^* = \underbrace{\bar{\mathbf{H}}_t}_{\text{scene-contact refined}} + \underbrace{\Delta\mathbf{H}_t^{ho}}_{\text{object-interaction residual}}, \qquad \mathbf{Y}_t = \mathrm{Head}_{\mathrm{human}}(\mathbf{H}_t^*). \tag{6}$$

The object residual updates both the object-pose and human-object contact representations, yielding $\mathbf{o}_t = \mathbf{o}_t^{\mathrm{pre}} + \Delta\mathbf{o}_t$ and $\mathbf{k}_t = \mathbf{k}_t^{\mathrm{pre}} + \Delta\mathbf{o}_t$. The construction of the pre-interaction representations $\mathbf{o}_t^{\mathrm{pre}}$ and $\mathbf{k}_t^{\mathrm{pre}}$, together with the interaction details, is given in the **Sup. Mat.**.

These representations predict dense human-object contact and a human-relative object pose:

$$\hat{\mathbf{c}}_t^{ho} = \sigma\Big(\mathrm{Head}_{\mathrm{contact}}^{ho}(\mathbf{k}_t)\Big), \qquad (\bar{R}_t, \bar{\mathbf{t}}_t) = \mathrm{Head}_{\mathrm{pose}}(\mathbf{o}_t). \tag{7}$$

The pose head predicts the object relative to the reconstructed human. Using the predicted human root rotation R_t^h and camera-frame pelvis position $\mathbf{p}_t^h$, we obtain the initial camera-frame object pose as $R_t^0 = R_t^h\bar{R}_t$ and $\mathbf{t}_t^0 = R_t^h\bar{\mathbf{t}}_t + \mathbf{p}_t^h$. This interaction-conditioned pose provides the initialization for the mesh-indexed object pose readout.

4.2 Mesh-Indexed Object Pose Readout

The initial pose guides where we gather visual evidence for points on the known object template. For each of Q canonical template points $\{\mathbf{q}_i\}_{i=1}^Q$, we project its initial camera-space location $\mathbf{y}_{t,i}^0 = R_t^0\mathbf{q}_i + \mathbf{t}_t^0$ into the object crop. Since this interaction-conditioned pose is only an initial estimate, the projected template points may still be misaligned with the true object appearance. The query feature $\widetilde{\mathbf{Q}}_{t,i}$ therefore combines the locally sampled image feature, canonical point geometry, global crop context, and interaction-conditioned object representation $\mathbf{o}_t$ to correct each projected point.

A regression head predicts a camera-space point correction. The corrected points define the correspondences for a single rigid pose, recovered by closed-form Kabsch alignment (Kabsch, 1978):

$$(R_t^o, \mathbf{t}_t^o) = \mathrm{Kabsch}\left(\{\mathbf{q}_i\}_{i=1}^Q, \{\hat{\mathbf{y}}_{t,i}\}_{i=1}^Q\right) \quad \text{where,} \quad \hat{\mathbf{y}}_{t,i} = \mathbf{y}_{t,i}^0 + \mathrm{MLP}_{\Delta}(\widetilde{\mathbf{Q}}_{t,i}). \tag{8}$$

Applying this transform to the full metric template preserves its shape and scale without iterative pose optimization at inference.

4.3 Training

Training follows the two components of the method. We first train the interaction-aware initialization (Section 4.1) together with the inherited human-scene reconstruction to establish the coupled representation and initial object pose. We then freeze this complete pathway and train only the mesh-query correction head (Section 4.2) to refine the pose using object-specific visual evidence.

Interaction-aware initialization. The first-stage objective is $\mathcal{L}_{\mathrm{init}} = \mathcal{L}_{\mathrm{HS}} + \mathcal{L}_{\mathrm{HO}} + \mathcal{L}_{\mathrm{obj}} + \mathcal{L}_{\mathrm{aux}}$, where $\mathcal{L}_{\mathrm{HS}}$ contains the inherited human-scene losses, $\mathcal{L}_{\mathrm{HO}}$ supervises human-object contact, and $\mathcal{L}_{\mathrm{obj}}$ supervises the human-relative object pose, and $\mathcal{L}_{\mathrm{aux}}$ is the auxiliary supervision loss.

Mesh-indexed object pose learning. We supervise the recovered rigid pose and transformed template, rather than individual query corrections, using $\mathcal{L}_{\mathrm{mesh}} = \mathcal{L}_{\mathrm{pose}} + \mathcal{L}_{\mathrm{geom}}$, where $\mathcal{L}_{\mathrm{pose}}$ supervises rotation and translation, and $\mathcal{L}_{\mathrm{geom}}$ supervises template geometry. Symmetry-equivalent poses are treated as equivalent targets. Details on loss functions are mentioned in the **Sup. Mat.**.

Table 1: **Global human motion estimation on EMDB-2 and RICH.** Human3R* denotes the released checkpoint evaluated in our pipeline, while Human3R† denotes the same architecture trained with the UniCon3R protocol. WA-MPJPE, W-MPJPE, and PA-MPJPE are reported in mm; RTE is reported in %. Lower is better for all metrics. **Bold**: best and underlined: second-best.

Method	EMDB-2 (Kaufmann et al., 2023)				RICH Huang et al. (2022a)			
	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	PA-MPJPE ↓
JOSH3R Liu et al. (2026)	220.0	661.7	13.1	–	190.4	334.9	6.3	–
UniSH Li et al. (2026)	118.5	270.1	5.8	44.4	118.1	183.2	4.8	50.0
Human3R* Chen et al. (2026)	118.2	286.1	2.4	38.7	109.9	184.0	3.3	48.6
Human3R† Chen et al. (2026)	156.7	378.6	3.5	40.6	97.5	153.2	2.3	41.4
UniCon3R Sur et al. (2026)	113.7	285.6	2.3	38.2	81.5	129.8	**1.7**	37.6
METAV3RSE (Ours)	**87.0**	**220.6**	**2.2**	**34.5**	**79.1**	**126.2**	**1.7**	**37.1**

5 EXPERIMENTS

We organize the experiments as follows. We first evaluate global human motion and local human mesh recovery (Section 5.1), beginning with the held-out EMDB benchmarks that are not used for training. We then evaluate human-object reconstruction and contact on InterCap and BEHAVE (Section 5.2). Finally, the ablation study (Section 5.3) evaluates the design choices of the mesh-indexed pose refinement. Our primary comparison is with feed-forward methods, since METAV3RSE performs direct inference without per-sequence test-time optimization.

Datasets. Following UniCon3R (Sur et al., 2026), global human motion is evaluated on EMDB-2 (Kaufmann et al., 2023) and moving-camera RICH (Huang et al., 2022a), while local human mesh recovery is evaluated on 3DPW (von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023). For human-object reconstruction and contact, we evaluate on InterCap (Huang et al., 2022c) and BEHAVE (Bhatnagar et al., 2022). We follow the MILO evaluation list (Chatterjee & Pavlakos, 2026) for InterCap and the VisTracker split (Xie et al., 2023) for BEHAVE. The human-scene pathway is trained on RICH and BEDLAM, while object and interaction supervision uses BEHAVE, InterCap, and HODome. Importantly, neither EMDB nor 3DPW is used for training, making EMDB a held-out out-of-distribution evaluation of the human reconstruction pathway.

Metrics. For global human motion, we report WA-MPJPE, W-MPJPE, and RTE. For local human mesh recovery, we report PVE, MPJPE, and PA-MPJPE. For human-object reconstruction, we report human, object, and combined Chamfer distance (CD_h, CD_o, CD_c), object vertex-to-vertex error ($\mathrm{V2V}_o$), and, on BEHAVE, human and object acceleration error (Acc_h, Acc_o). Human-object contact is evaluated using F1. All reconstruction and motion errors are lower-is-better except F1.

Baselines. For human reconstruction, we follow the baselines and evaluation protocol of Human3R and UniCon3R (Chen et al., 2026; Sur et al., 2026). Our main human-object baselines are the feed-forward monocular methods HOI-TG (Wang et al., 2025b) and CONTHO (Nam et al., 2024), evaluated on the same frames using the same dataset-specific scoring procedure. Both operate on individual images. METAV3RSE uses independent-frame inference on InterCap and recurrent video inference on BEHAVE.

Optimization-based methods such as VisTracker (Xie et al., 2023) and CARI4D (Xie et al., 2026) require test-time fitting and therefore follow a different inference regime. THO (Zhang et al., 2026) is feed-forward, but a compatible public checkpoint is unavailable for our evaluation protocol. For human-object reconstruction on BEHAVE, we follow the released VisTracker protocol (Xie et al., 2023) used in our main evaluation. We additionally report results under the first-frame alignment protocol described by THO in the **Sup.Mat.**, using the same predictions and scoring pipeline.

For human-object evaluation, METAV3RSE uses calibrated intrinsics, a known metric object template, and supplied object localization, while human localization is predicted. HOI-TG and CONTHO use human/object segmentations and annotation crops, and HOI-TG additionally uses known object geometry.

Table 2: **Local human mesh recovery on 3DPW and EMDB-1.** Human3R* denotes the released checkpoint evaluated in our pipeline. PVE denotes human per-vertex error. PVE, MPJPE, and PA-MPJPE are reported in mm; lower is better for all metrics. **Bold**: best and underlined: second-best.

Method	**3DPW** von Marcard et al. (2018)			**EMDB-1** (Kaufmann et al., 2023)		
	PVE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓
UniSH Li et al. (2026)	94.2	80.1	–	–	–	–
Human3R* Chen et al. (2026)	86.6	72.8	44.6	86.8	74.6	49.2
UniCon3R Sur et al. (2026)	86.8	72.5	45.4	94.3	81.7	49.8
METAV3RSE (Ours)	**83.1**	**72.0**	**42.5**	**75.6**	**65.7**	**45.0**

Table 3: **Human–object reconstruction and contact on InterCap and BEHAVE.** Input lists InterCap/BEHAVE context (I: single image; V: video). Reconstruction distances are reported in cm and acceleration errors in cm/frame^2; lower is better. Human-object contact is evaluated using F1.

Method	Input	**InterCap** (Huang et al., 2022c)					**BEHAVE** (Bhatnagar et al., 2022)						
		CD_h ↓	CD_o ↓	CD_c ↓	$V2V_o$ ↓	F1 ↑	CD_h ↓	CD_o ↓	CD_c ↓	$V2V_o$ ↓	Acc_h ↓	Acc_o ↓	F1 ↑
HOI-TG (Wang et al., 2025b)	I/I	8.7	**26.9**	9.7	34.5	0.56	34.0	22.4	24.6	**25.6**	**1.7**	**5.3**	0.60
CONTHO (Nam et al., 2024)	I/I	8.9	27.8	10.0	35.2	**0.63**	34.3	22.5	24.8	25.7	1.9	5.5	**0.66**
METAV3RSE (Ours)	I/V	**7.4**	**26.9**	**8.7**	**32.9**	0.50	**22.5**	**21.6**	**18.8**	28.4	2.8	7.3	0.65

5.1 GLOBAL HUMAN MOTION AND LOCAL HUMAN MESH RECONSTRUCTION

Global human motion. We first evaluate on EMDB-2, which is not used during training and therefore serves as an out-of-distribution benchmark for the recovered human motion. As shown in Table 1, METAV3RSE reduces WA-MPJPE from 113.7 to 87.0 mm and W-MPJPE from 285.6 to 220.6 mm relative to UniCon3R, improvements of 23.5% and 22.8%, respectively. RTE decreases from 2.3% to 2.2%, while PA-MPJPE improves from 38.2 to 34.5 mm. These gains are particularly notable because EMDB is entirely held out from training.

On RICH (Table 1), METAV3RSE achieves the best or tied-best results among the compared methods, with modest gains over UniCon3R on all three pose metrics and unchanged RTE.

Local human mesh recovery. On 3DPW (Table 2), METAV3RSE achieves the lowest PVE, MPJPE, and PA-MPJPE among the listed methods. On EMDB-1, which is likewise not used during training, PVE, MPJPE, and PA-MPJPE improve by 11.2, 8.9, and 4.2 mm, respectively, over the strongest listed baseline for each metric. Together with the EMDB-2 global-motion results, this indicates that the improvement in human reconstruction extends beyond the training distribution.

5.2 4D HUMAN-OBJECT RECONSTRUCTION

Human-object reconstruction. On InterCap (Table 3), where all methods are evaluated with independent-frame inference, METAV3RSE obtains the lowest human CD, combined CD, and object V2V, while tying the best object CD among the compared methods. Relative to HOI-TG, combined CD decreases by 10.3%. These gains therefore persist even without recurrent video context.

On BEHAVE, METAV3RSE achieves the lowest human, object, and combined Chamfer distances among the compared methods. Relative to HOI-TG, human and combined errors decrease by 33.8% and 23.6%, while object CD improves from 22.4 to 21.6 cm. Fig. 2 illustrates the corresponding failure modes. HOI-TG and CONTHO frequently exhibit hand-object misregistration, where the reconstructed hand is displaced from the object or fails to maintain the intended grasp. In contrast, METAV3RSE has more consistent hand-object alignments, including under partial object occlusion.

The remaining BEHAVE object metrics are less uniform. Object V2V and both acceleration errors are higher than those of HOI-TG and CONTHO. Lower surface error therefore does not imply better correspondence-sensitive pose or temporal accuracy; Stage 2 adds no temporal objective. The InterCap examples in Fig. 2 further show incorrect relative depth or orientation in the baseline reconstructions, leading to human-object interpenetration for interactions such as the umbrella and ball. We preserve the relative human-object placement and the rigid object structure.

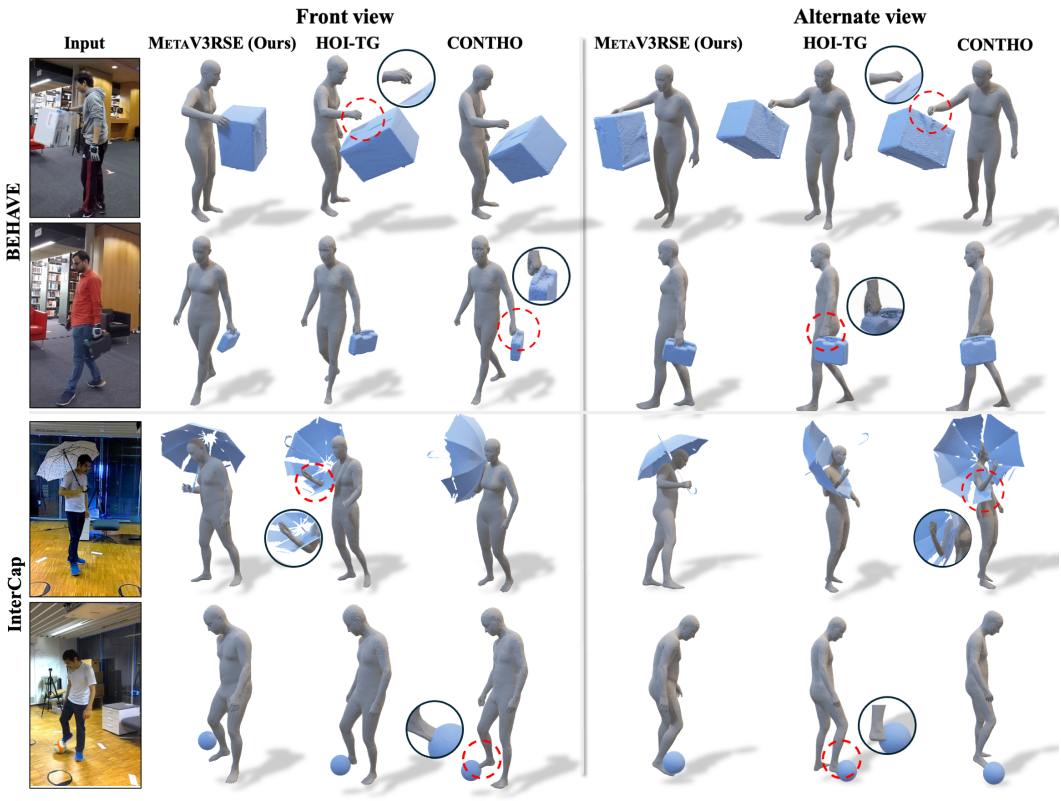


Figure 2: **Qualitative comparison of human-object reconstruction on InterCap and BEHAVE.** On InterCap, alternate views reveal incorrect relative placement and human-object interpenetration in the baselines. On BEHAVE, red circles highlight hand-object misregistration in HOI-TG and CONTHO, where the reconstructed hand is displaced from the object or fails to maintain the intended grasp. METAV3RSE better preserves the human-object configuration across interactions.

Human-object contact. On InterCap, contact F1 is lower than both baselines despite the improved combined reconstruction, indicating different rankings across the contact and reconstruction metrics. On BEHAVE, contact F1 is 0.65, versus CONTHO's 0.66. Fig. 3 shows recovery of the active hand contact under partial occlusion, where baselines miss it or predict contact elsewhere.

Overall, the results support two observations. First, interaction-aware reconstruction improves the recovered human-object configuration, particularly under object occlusion. Second, the remaining visible object evidence is useful for correcting the rigid pose beyond the interaction-conditioned initialization. Additional human-object contact comparisons on InterCap and BEHAVE, and human-scene contact results on RICH, are provided in the **Sup. Mat.**

5.3 ABLATION STUDIES

Rigid pose readout. As observed in Table 4, compared with direct pose regression, the mesh-indexed readout reduces object CD by 0.7 cm on BEHAVE and 3.0 cm on InterCap. Object V2V improves on BEHAVE but is slightly worse on InterCap. Thus, predicting point evidence before recovering the rigid pose improves object surface alignment on both datasets.

Query conditioning. Removing local crop features increases object CD by 1.3 cm on BEHAVE and 5.0 cm on InterCap. This confirms that local visual evidence contributes to object-pose refinement beyond the global crop and interaction context. Removing the interaction conditioning worsens all reported object metrics on both datasets. Since initialization is fixed, this isolates the benefit of retaining interaction context in the readout. Removing global crop attention increases object CD by 0.9 cm on BEHAVE and 3.7 cm on InterCap, but slightly reduces BEHAVE acceleration error (7.2 vs. 7.3). Its benefit is therefore primarily spatial.

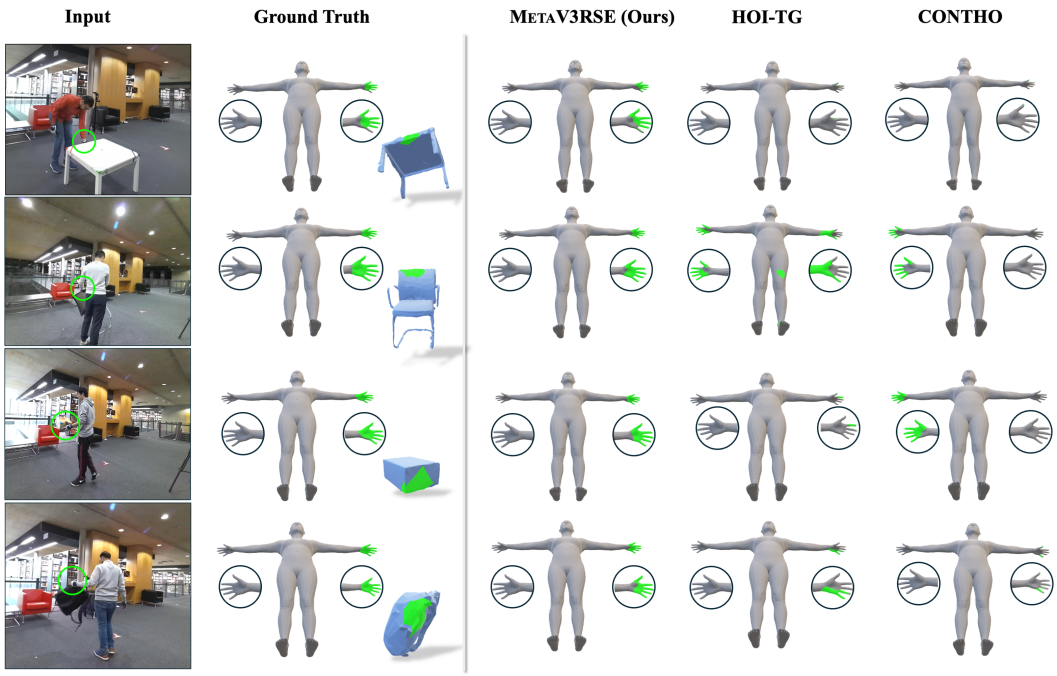


Figure 3: **Qualitative comparison of human-object contact on BEHAVE.** Green vertices indicate human-object contact. Ground-truth contact lies primarily on the active right hand. METAV3RSE recovers this region even under partial occlusion or weak visibility, whereas the baselines may miss the interacting hand or predict contact on non-interacting body regions. Circular insets enlarge the corresponding hand predictions.

Table 4: **Ablation study of the mesh-indexed object pose readout.** We evaluate the rigid pose readout and the visual and interaction cues used to condition the mesh queries on BEHAVE and InterCap. All variants use the same interaction-aware object initialization. CD and V2V are reported in cm and acceleration error in cm/frame^2; lower is better for all metrics.

	BEHAVE (Bhatnagar et al., 2022)			**InterCap** (Huang et al., 2022c)	
Variant	$CD_o \downarrow$	$V2V_o \downarrow$	$Acc_o \downarrow$	$CD_o \downarrow$	$V2V_o \downarrow$
Direct pose regression	<u>22.3</u>	<u>28.8</u>	7.4	<u>29.9</u>	**32.5**
w/o Local Crop Features	22.9	29.1	7.9	31.9	34.6
w/o Global Crop Attention	22.5	28.9	**7.2**	30.6	33.5
w/o Interaction Conditioning	24.0	29.8	8.7	30.7	33.5
METAV3RSE (Ours)	**21.6**	**28.4**	<u>7.3</u>	**26.9**	<u>32.9</u>

6 CONCLUSION

We presented METAV3RSE, a unified feed-forward framework for 4D human-object-scene reconstruction from monocular video. Starting from a scene-grounded human reconstruction, METAV3RSE introduces human-object interaction directly into the reconstruction pathway: the reconstructed human provides an interaction-conditioned object-pose hypothesis, object interaction feeds back to refine the human, and a mesh-indexed readout uses the resulting pose to determine where object-specific visual evidence should be gathered. The final object pose is recovered through a closed-form rigid readout, without per-frame or per-sequence pose optimization.

Across our experiments, a consistent picture emerges. METAV3RSE suggests a broader direction for unified 4D perception: humans, manipulated objects, and scenes should not be reconstructed independently and reconciled afterwards. Their interactions should instead determine how information is exchanged and how each component is corrected inside the reconstruction model itself.

AI USE STATEMENT

We used generative AI tools only for language editing and presentation of the manuscript, including improving clarity and concision, refining section organization, and drafting alternative formulations of text, captions, and figure layouts.

Generative AI was not used to generate research ideas, model architectures, experimental procedures, source code, benchmark labels, ground-truth annotations, or reported quantitative results. All model design choices, experiments, evaluation procedures, numerical results, and scientific conclusions were developed and verified by the authors. AI-assisted text and figure suggestions were manually checked against the implemented method, experiment logs, and evaluated model before inclusion in the paper. Literature-related statements and comparisons were likewise checked against the corresponding papers by the authors.

ETHICS STATEMENT

This work uses existing research datasets and does not introduce a new human-subject data collection. We use datasets including BEHAVE (Bhatnagar et al., 2022), InterCap (Huang et al., 2022c), HODome (Zhang et al., 2023), RICH (Huang et al., 2022a), BEDLAM (Black et al., 2023), 3DPW (von Marcard et al., 2018), and EMDB, following their respective research-use terms and evaluation protocols.

Because our method reconstructs human motion, human-object interaction, and scene geometry from monocular video, similar technology could in principle be applied to videos containing people without their knowledge or consent. Such use raises privacy and surveillance concerns. Our work is intended for research on 3D/4D perception and embodied interaction and does not perform identity recognition or infer sensitive personal attributes.

As with other learning-based human reconstruction systems, performance may depend on the distributions represented in the training data, including body appearance, clothing, motion, object categories, occlusion patterns, and capture conditions. Reconstruction accuracy may therefore degrade for people, interactions, or environments that are under-represented in the training data. These limitations should be considered when applying the method outside the research settings evaluated in this paper.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the proposed method and evaluation across the main paper and supplementary material. The methodology describes the interaction-aware object initialization, human-object refinement, mesh-indexed object pose readout, and sequential training procedure. The implementation details specify the backbone configuration, prompt and feature construction, patch selection, interaction iterations, template-query sampling, projection procedure, rigid Kabsch readout, loss definitions, symmetry handling, optimization stages, and training data.

We additionally document the evaluation datasets, metrics, object-localization assumptions, and benchmark protocols used for BEHAVE, InterCap, global human motion, local human mesh reconstruction, and contact evaluation. The supplementary material further reports backbone scaling, first-frame alignment, object-pose analysis, and relevant ablations. Code and checkpoints will be released to support reproducibility of the reported results.

REFERENCES

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

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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*, 2023.

Agniv Chatterjee and Georgios Pavlakos. Reconstructing humans and objects in interaction using large reconstruction models. In *European Conference on Computer Vision*, 2026.

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

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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*, 2023.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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*, 2022a.

Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *European Conference on Computer Vision*, pp. 585–603. Springer, 2022b.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, volume 13485. Springer, 2022c.

Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 34(5):827–828, 1978.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

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

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Ao Li, Jinpeng Liu, Yixuan Zhu, and Yansong Tang. Scorehoi: Physically plausible reconstruction of human-object interaction via score-guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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

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

Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. HOSNeRF: Dynamic human-object-scene neural radiance fields from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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

Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3D human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

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

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tanuj Sur, Shashank Tripathi, Nikos Athanasiou, Ha Linh Nguyen, Kai Xu, Michael J. Black, and Angela Yao. UniCon3R: Unified contact-aware 4D human-scene reconstruction from monocular video. *arXiv preprint arXiv:2604.19923*, 2026.

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*, 2023.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Zhenrong Wang, Qi Zheng, Sihan Ma, Maosheng Ye, Yibing Zhan, and Dongjiang Li. End-to-end HOI reconstruction transformer with graph-based encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single RGB image. In *European Conference on Computer Vision*. Springer, 2022.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. CARI4D: Category agnostic 4D reconstruction of human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Lixin Xue, Chengwei Zheng, Georgios Paschalidis, Chen Guo, Manuel Kaufmann, Juan Zarate, and Dimitrios Tzionas. RHINO: Reconstructing human interactions with novel objects from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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*, pp. 6069–6079. IEEE, 2025.

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*, pp. 13846–13855, 2026.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Haoyu Zhang, Wei Zhai, Yuhang Yang, Yang Cao, and Zheng-Jun Zha. End-to-end spatial-temporal transformer for real-time 4D HOI reconstruction. *arXiv preprint arXiv:2603.14435*, 2026.

Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision*. Springer, 2020.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Appendix

Table of Contents

A Implementation Details 14

B Analysis of the Mesh-Indexed Object Pose Readout 18
B.1 Object Pose Readout under Occlusion 18
B.2 Disentangling Object Placement and Orientation 19

C Evaluation under First-Frame Alignment 19

D Backbone Scaling 20

E Contact Evaluation 22

F Failure Analysis 23

G Additional Qualitative Results 23

A Implementation Details

Notation and inherited backbone. METAV3RSE builds on the recurrent human-scene reconstruction of UniCon3R (Sur et al., 2026), which in turn inherits the Human3R human-reconstruction pathway. The recurrent decoder has embedding dimension 768. The image encoder, recurrent scene backbone, and scene/camera heads are inherited from the pretrained reconstruction model and remain frozen. Human-prior features are extracted using the frozen Multi-HMR encoder and sampled at the predicted human anchor.

Throughout this section, blue denotes modules inherited from the frozen reconstruction backbone, while red denotes task-specific prediction or interaction modules introduced or adapted beyond that backbone. This color convention describes the architectural origin of a module rather than its trainability in every stage: during mesh-indexed training, the complete interaction-aware initializer is frozen and only the mesh-query correction head is optimized. We use LN for layer normalization, CA for multi-head cross-attention, and $\oplus$ for channel-wise concatenation.

Human and object localization. The local interaction module should operate only on image regions belonging to the interacting human and object rather than on the complete image feature grid. We therefore use the predicted human mask $\mathbf{M}_t^h$ from the inherited human detection/segmentation pathway and the supplied object mask $\mathbf{M}_t^o$ to select human and object regions from the decoded image features $\mathbf{F}'_t$.

The masks act only as spatial selectors: they do not replace or modify the feature channels themselves. If $\mathbf{F}'_t$ contains N patch features of dimension D, mask-based selection yields $\mathbf{F}_t^h \in \mathbb{R}^{N_h \times D}$ and $\mathbf{F}_t^o \in \mathbb{R}^{N_o \times D}$, where N_h and N_o are the numbers of retained human and object patches. We

retain at most 64 patches per entity according to mask coverage and discard patches whose entity coverage is below 0.25.

Human-scene and human-object contact are both represented on the 6,890-vertex SMPL body topology. The initial object pose is represented relative to the reconstructed human using the continuous 6D rotation representation of Zhou et al. (2019); it is later composed with the predicted human root orientation and pelvis position as described in Section 4.1.

Why separate object-pose and contact representations? The recurrent decoder produces the decoded object prompt $\mathbf{O}'_t$ and human-object contact prompt $\mathbf{C}'^{ho}_t$ in the same recurrent context as the human and scene. These representations provide shared interaction context, but the two downstream tasks require different information. Object-pose estimation needs strong evidence about the appearance and location of the manipulated object, whereas contact prediction needs a representation of the relation between the human and object. We therefore construct separate pre-interaction features for these two purposes.

Object-pose feature construction. Although $\mathbf{O}'_t$ already carries object context, the pose readout benefits from a dedicated query that can retrieve image evidence specifically for object localization. We initialize this query from the decoded human representation,

$$\mathbf{q}^o_t = \mathbf{H}'_t + \mathbf{e}_{\text{iso}}, \tag{9}$$

where $\mathbf{e}_{\text{iso}}$ is a learned embedding that specializes the human-conditioned query for object-pose estimation. This is distinct from the recurrent object prompt $\mathbf{O}_t$: $\mathbf{O}_t$ inserts object context into the shared recurrent decoder, whereas $\mathbf{q}^o_t$ is used after decoding to construct the feature read by the object-pose head.

The query retrieves object-specific evidence from the decoded image features:

$$\mathbf{r}_t = \mathrm{LN}(\mathbf{q}^o_t + \mathrm{CA}_{\text{iso}}(\mathbf{q}^o_t, \mathbf{F}'_t)). \tag{10}$$

Here, CA_{iso} is a dedicated cross-attention layer for the object-pose readout. The resulting feature $\mathbf{r}_t$ therefore combines the reconstructed-human reference with image evidence relevant to the object.

We further refine this feature using $\mathrm{MLP}_{\text{iso}}$, a learned residual MLP dedicated to the object-pose representation, and concatenate the human prior:

$$\mathbf{o}^{\text{pre}}_t = [\mathbf{r}_t + \mathrm{MLP}_{\text{iso}}(\mathrm{LN}(\mathbf{r}_t))] \oplus \mathbf{F}_{\text{HMR},t}. \tag{11}$$

We call this $\mathbf{o}^{\text{pre}}_t$ because it is the object-pose representation *before* explicit local human-object interaction is incorporated. The human prior is useful here because the object pose is predicted relative to the reconstructed body.

Human-object contact feature construction. Contact requires a different feature. Rather than asking only where the object appears, the contact branch should explicitly relate the decoded contact representation to the decoded object representation. We therefore let $\mathbf{C}'^{ho}_t$ retrieve object context from $\mathbf{O}'_t$:

$$\mathbf{k}^{\text{pre}}_t = \left[\mathbf{C}'^{ho}_t + \mathrm{CA}_{ho}(\mathbf{C}'^{ho}_t, \mathbf{O}'_t)\right] \oplus \mathbf{F}_{\text{HMR},t}. \tag{12}$$

Here, CA_{ho} is a dedicated cross-attention layer for the human-object contact representation. Thus, $\mathbf{k}^{\text{pre}}_t$ is the human-object contact representation before the explicit local patch interaction. In contrast to $\mathbf{o}^{\text{pre}}_t$, which emphasizes object-specific image evidence, $\mathbf{k}^{\text{pre}}_t$ emphasizes the relation between decoded object and contact context.

Human-object interaction module. The representations above capture global interaction context, but they do not explicitly determine which local human regions interact with which local object regions. We therefore apply the $\mathrm{Interact}$ operator from Section 4.1 to the mask-selected patch features $\mathbf{F}^h_t$ and $\mathbf{F}^o_t$.

Each valid patch is projected into a 128-dimensional pair space. Learned affinities are then computed between human and object patches. The interaction is bidirectional: a human patch aggregates information from related object patches, while an object patch aggregates information from related human patches. Separate residual mappings convert these aggregated features into human-side and object-side updates.

Because these two sets of patch features condition one another, we refine the interaction through a small number of damped iterations rather than using a single exchange. Starting from $\mathbf{Z}^0 = 0$,

$$\mathbf{Z}^{k+1} = \frac{1}{2}\mathbf{Z}^k + \frac{1}{2}g\big(\mathbf{Z}^k; \mathbf{F}_t^h, \mathbf{F}_t^o\big)\,, \tag{13}$$

where g denotes one bidirectional patch-interaction update. At every iteration it recomputes human-object affinities, aggregates features from the opposite entity, and predicts updated patch residuals. We run at most eight iterations and stop when the residual falls below 10^{-4}.

The final valid patch updates are pooled separately over the two entities, giving the human residual $\Delta\mathbf{H}_t^{ho}$ and object-side residual $\Delta\mathbf{o}_t$. The human residual is added after the inherited human-scene correction and before SMPL-X regression, as described in Eq. (6).

The object-side residual updates both downstream representations:

$$\mathbf{o}_t = \mathbf{o}_t^{\mathrm{pre}} + \Delta\mathbf{o}_t, \qquad \mathbf{k}_t = \mathbf{k}_t^{\mathrm{pre}} + \Delta\mathbf{o}_t. \tag{14}$$

Thus, $\mathbf{o}_t$ is the final interaction-conditioned representation used for object-pose prediction, while $\mathbf{k}_t$ is the final interaction-conditioned representation used for human-object contact prediction. These are the two representations decoded in Eq. (7).

Mesh-indexed object pose readout. The interaction-aware stage provides an approximate rigid object pose, but the pose alone does not tell the model how visible image evidence corresponds to specific locations on the known object. The mesh-indexed readout makes this association explicit.

We sample $Q = 64$ canonical points from the metric object template using deterministic farthest-point sampling. These points act as persistent geometric identities: each query always refers to the same canonical location on the object. They are distinct from the denser template vertices used only for the training losses.

Under the initial object pose, each canonical point is transformed into camera space and projected into the supplied object crop using the calibrated source intrinsics and crop box. The crop is resized to 224×224 and encoded by a frozen image encoder into a 14×14 grid of 1,024-dimensional image features.

For each canonical query, we combine four complementary sources of evidence: (i) its canonical 3D position, which identifies where it lies on the template; (ii) a local crop feature bilinearly sampled at its projected image location; (iii) global crop context aggregated over the complete crop feature grid; and (iv) the interaction-conditioned object representation $\mathbf{o}_t$, which carries the human-object context established by the first stage.

The resulting query representation therefore answers a specific question: given where the interaction-aware initializer expects this template point to appear, what do the local image evidence, the complete object appearance, and the human-object interaction imply about its actual camera-space location?

The mesh-query module uses hidden dimension 256 with eight attention heads. Its output MLP predicts the camera-space correction described in Section 4.2. Queries projected behind the camera or outside the object crop receive no local crop feature, although their canonical, global-crop, and interaction-conditioned context remains available. The spatial readout does not estimate camera intrinsics and does not use ground-truth human or object pose at inference.

Closed-form rigid readout. The corrected query locations are not interpreted as a deformable reconstruction of the object. Their purpose is to provide a set of correspondences between fixed canonical template locations and their predicted camera-space positions. Since the object template is rigid, these correspondences should be explained by one rotation and translation.

We therefore recover the final object pose using determinant-corrected Kabsch alignment (Kabsch, 1978) between the canonical query points and their predicted camera-space locations. The resulting rigid transformation is applied to the complete metric template. The Kabsch readout contains no learned parameters and requires no iterative object-pose optimization at inference.

Interaction-initialization supervision. The first training stage learns the representations needed to produce the scene-grounded human, human-object contact, and initial human-relative object pose.

It uses the interaction-aware initialization objective defined in Section 4.3. The inherited term $\mathcal{L}_{\mathrm{HS}}$ contains the human-scene reconstruction, SMPL-X, and human-scene contact objectives inherited from UniCon3R. $\mathcal{L}_{\mathrm{HO}}$ supervises human-object contact on the human-body vertices, while $\mathcal{L}_{\mathrm{obj}}$ supervises the relative object rotation and translation.

The auxiliary term $\mathcal{L}_{\mathrm{aux}}$ is used to structure the internal interaction representations during this stage. An auxiliary dual-pointmap branch predicts corresponding camera-frame and canonical object coordinates together with their validity, while patch-pair supervision provides targets for the internal human-object affinities. These auxiliary outputs support learning of the interaction-aware initializer but do not determine the final object pose used at inference. Each term is applied only when the corresponding supervision is available.

For the damped interaction module, gradients are computed through the final application of g after the iterative state has been obtained, rather than backpropagating through the complete sequence of interaction iterations.

Mesh-indexed object supervision. The second stage has a narrower purpose: given the fixed interaction-conditioned pose hypothesis, it learns how template-indexed visual evidence should correct that hypothesis. We therefore freeze the complete human, scene, contact, interaction, and initial object-pose pathways in evaluation mode and optimize only the mesh-query correction head. Supervision is applied after the rigid Kabsch readout rather than directly to individual query offsets, because the query predictions are only intermediate evidence for one rigid object pose.

Object symmetries may admit multiple equivalent ground-truth poses. During training, we simply evaluate the loss against each pre-defined symmetry candidate and retain the lowest-loss target. This is a finite target-selection step used only for supervision; it does not optimize the predicted object pose and introduces no iterative procedure at inference.

For a ground-truth pose $(R^*, \mathbf{t}^*)$ and a valid canonical symmetry $(R_s, \mathbf{t}_s)$, the equivalent target pose is

$$R_s^* = R^* R_s, \qquad \mathbf{t}_s^* = \mathbf{t}^* + R^* \mathbf{t}_s. \tag{15}$$

Spatial supervision uses at most $M = 256$ template vertices, sampled independently of the 64 mesh queries. This separation is intentional: the queries determine the predicted pose, whereas the denser vertices measure how well that rigid pose transforms the object geometry. For a predicted pose $(R, \mathbf{t})$ and symmetry-equivalent target $(R_s^*, \mathbf{t}_s^*)$, let

$$\mathbf{p}_j = R\mathbf{v}_j + \mathbf{t}, \qquad \mathbf{p}_{j,s}^* = R_s^* \mathbf{v}_j + \mathbf{t}_s^*. \tag{16}$$

We supervise four complementary aspects of the recovered rigid pose: rotation, translation, corresponding posed-template points, and overall surface agreement:

$$\begin{aligned}
\ell_R^s &= \frac{1}{2} \left\| R - R_s^* \right\|_F^2, \qquad \ell_t^s = \frac{1}{3} \left\| \mathbf{t} - \mathbf{t}_s^* \right\|_1, \\
\ell_p^s &= \frac{1}{3M} \sum_{j=1}^{M} \left\| \mathbf{p}_j - \mathbf{p}_{j,s}^* \right\|_1, \\
\ell_d^s &= \frac{1}{M} \sum_j \min_k \left\| \mathbf{p}_j - \mathbf{p}_{k,s}^* \right\|_2 + \frac{1}{M} \sum_k \min_j \left\| \mathbf{p}_j - \mathbf{p}_{k,s}^* \right\|_2.
\end{aligned} \tag{17}$$

The rotation and translation terms directly supervise the recovered rigid transform. The corresponding-point term preserves the canonical correspondence implied by the template, while the bidirectional nearest-neighbor term measures overall surface agreement and is less sensitive to a particular vertex indexing.

The final spatial objective is

$$\mathcal{L}_{\mathrm{mesh}} = \min_{s \in \mathcal{S}} \left(0.1\, \ell_R^s + \ell_t^s + \ell_p^s + \ell_d^s \right). \tag{18}$$

The minimization above only selects among a finite set of symmetry-equivalent training targets. At inference, the object pose is obtained directly from the network predictions followed by the closed-form Kabsch readout; no pose variables are iteratively optimized.

Table 5: **Mesh-indexed object pose readout under body occlusion.** Object Chamfer distance CD_o is reported in cm on InterCap (Huang et al., 2022c). For this diagnostic, a similarity transform estimated from the predicted and ground-truth human meshes is applied identically to the initialization and final object predictions; the object is not aligned independently. $\Delta\mathrm{CD}_o$ denotes the reduction from the interaction-conditioned initialization to the final mesh-indexed readout.

Occluded object	Images/captures	Init. $\mathrm{CD}_o\downarrow$	Final $\mathrm{CD}_o\downarrow$	$\Delta\mathrm{CD}_o\uparrow$
0–25%	1,163/34	76.02	45.04	30.98
25–50%	119/25	56.11	43.17	12.93
50–75%	82/25	66.21	40.00	26.22
75–100%	32/10	88.01	74.43	13.58
All images	**1,396/–**	**74.02**	**45.26**	**28.76**

This objective corresponds to $\mathcal{L}_{\mathrm{pose}} = 0.1\,\ell_R^s + \ell_t^s$ and $\mathcal{L}_{\mathrm{geom}} = \ell_p^s + \ell_d^s$ in Section 4.3. Translation and corresponding-point losses use mean coordinate-wise ℓ_1 error, while ℓ_d^s is the sum of the two directional mean nearest-neighbor distances. No additional temporal objective is introduced during this stage.

Training data and schedule. Object-bearing training slots comprise HODome (Zhang et al., 2023), BEHAVE-Date01, InterCap (Huang et al., 2022c), and BEHAVE (Bhatnagar et al., 2022). The training configuration also contains RICH (Huang et al., 2022a) and BEDLAM (Black et al., 2023); observations without valid object targets contribute zero spatial data loss.

The interaction-aware initializer is trained for 15 epochs. Starting from this initializer, the mesh-indexed stage is trained for 15 spatial epochs (9,690 optimizer updates), and this checkpoint is used for evaluation. The optimizer state is resumed from the preceding spatial run under a 40-epoch learning-rate horizon.

B Analysis of the Mesh-Indexed Object Pose Readout

B.1 Object Pose Readout under Occlusion

We test whether the mesh-indexed object pose readout remains useful when the manipulated object is partly occluded by the human. Following visibility-based analyses in human-object reconstruction (Xie et al., 2023), we render the annotated object both alone and together with the annotated human. The occlusion ratio is defined as the fraction of object pixels hidden after adding the human, and we partition the 1,396 InterCap (Huang et al., 2022c) images into four intervals. "Images/captures" in Table 5 reports the number of evaluated images and distinct interaction captures in each interval.

This experiment compares the interaction-conditioned object initialization with the final mesh-indexed object pose readout under a fixed diagnostic alignment. Both predictions use the same supplied object localization, so differences cannot be attributed to different masks or crops. We estimate a similarity transform from the predicted and ground-truth human meshes and apply the same transform to both object predictions. The object is never aligned independently, thereby preserving errors in its placement relative to the human. This human-derived alignment differs from the main InterCap evaluation protocol; therefore, the absolute CD_o values reported here are not directly comparable to those in the main human-object reconstruction table. We report bidirectional object Chamfer distance CD_o in centimeters together with the reduction from the initial to the final prediction.

As shown in Table 5, the mesh-indexed readout reduces mean object error in every occlusion interval. The all-images aggregate under this diagnostic protocol decreases from 74.02 cm to 45.26 cm. The largest absolute reduction is 30.98 cm for 0–25% occlusion, while the 50–75% interval improves by 26.22 cm. Even when more than 75% of the projected object is occluded, the error decreases from 88.01 cm to 74.43 cm. Relative to the initialization, the reductions are 40.7%, 23.0%, 39.6%, and 15.4% from the lowest to highest occlusion interval.

These results support the role of the two components in our object pathway. The interaction-conditioned initialization establishes an object-pose hypothesis from the reconstructed human-object

Table 6: **Object pose analysis on BEHAVE (Bhatnagar et al., 2022).** Centroid distance and ADD-S are reported in cm; ADD-S$<10\%d$ denotes the fraction of frames with ADD-S below 10% of the object diameter. *Prior* denotes the interaction-aware initialization from Section 4.1, while *final* denotes the output after the mesh-indexed readout of Section 4.2. Results are frame-weighted over all 127,655 evaluation frames. Best results are in **bold** and second-best results are underlined.

Method	Centroid ↓	ADD-S ↓	ADD-S$<10\%d$ ↑
HOI-TG (Wang et al., 2025b)	18.9	9.9	60%
CONTHO (Nam et al., 2024)	18.9	9.9	60%
METAV3RSE (Ours), prior	22.4	12.4	52%
METAV3RSE (Ours), final	**15.8**	**9.4**	**65%**

interaction, while the mesh-indexed readout uses the remaining object appearance to correct that hypothesis. The experiment does not suggest that severe occlusion is solved: the 75–100% interval contains only 32 images from 10 captures and retains the largest final error. We therefore treat this analysis as a within-model diagnostic of the object-pose readout rather than a separate cross-method benchmark.

B.2 DISENTANGLING OBJECT PLACEMENT AND ORIENTATION

Our object reconstruction proceeds in two steps. The Interaction-Aware Object Initialization (Section 4.1) predicts an object pose relative to the reconstructed human, yielding the initial camera-frame pose $(R_t^0, \mathbf{t}_t^0)$. The Mesh-Indexed Object Pose Readout (Section 4.2) then uses this pose to project canonical template points into the image, predict their camera-space corrections, and recover the final rigid pose through the closed-form readout in Eq. (8). We analyze what this second stage actually corrects.

The Chamfer metrics used in the main evaluation measure overall surface agreement, but do not distinguish errors in object placement from errors in orientation. We therefore additionally measure the distance between predicted and ground-truth object centroids and symmetry-aware ADD-S, and separately inspect the corresponding rotation error.

Table 6 shows that the mesh-indexed readout substantially corrects the initial object placement. Relative to the interaction-aware initialization, centroid error decreases from 22.4 to 15.8 cm and ADD-S from 12.4 to 9.4 cm, while the fraction of frames satisfying ADD-S$<10\%d$ increases from 52% to 65%. The final prediction also improves over HOI-TG and CONTHO on all three measures.

This behavior clarifies the role of the two object-pose components in METAV3RSE. The interaction-aware initialization uses the reconstructed human to establish an initial pose hypothesis, whereas the mesh-indexed stage uses the projected template locations to gather object-specific image evidence and correct that hypothesis. The centroid reduction is also observed directly in the camera frame, where the final readout improves the initialization by 15.8 cm, indicating that the gain is not introduced by the benchmark alignment. In contrast, rotation error changes little between the initial and final predictions. Empirically, the point corrections predicted by MLP_Δ in Eq. (8) therefore improve primarily the object's translational placement, while much of its orientation is already determined by the interaction-conditioned initialization.

C EVALUATION UNDER FIRST-FRAME ALIGNMENT

Our main BEHAVE (Bhatnagar et al., 2022) evaluation follows the released VisTracker protocol (Xie et al., 2023), which estimates a similarity alignment within each 300-frame evaluation window. We additionally evaluate under the first-frame alignment protocol described by THO (Zhang et al., 2026), where the alignment is estimated once from the first frame and kept fixed for the remainder of the sequence. In contrast to window-wise alignment, errors in the reconstructed coordinate frame can no longer be absorbed by repeatedly re-estimating the alignment. This provides a complementary evaluation of reconstruction consistency over longer temporal horizons. We rescore the same predictions used in the main evaluation, together with HOI-TG (Wang et al., 2025b) and CONTHO (Nam et al., 2024), on identical frames using the same scoring code.

Table 7: **BEHAVE (Bhatnagar et al., 2022) under first-frame alignment.** A single similarity transform is estimated from the first frame and kept fixed for the remainder of the sequence. Reconstruction distances are reported in cm; lower is better. All methods are evaluated on identical frames using the same scoring implementation. Best results are in **bold** and second-best results are underlined.

Method	$\mathrm{CD}_h \downarrow$	$\mathrm{CD}_o \downarrow$	$\mathrm{CD}_c \downarrow$	$\mathrm{V2V}_o \downarrow$
HOI-TG (Wang et al., 2025b)	<u>59.5</u>	62.0	51.5	51.8
CONTHO (Nam et al., 2024)	59.7	<u>60.8</u>	<u>51.3</u>	<u>51.3</u>
METAV3RSE (Ours)	**40.6**	**43.3**	**33.8**	**43.5**

Table 7 shows that the differences observed in the main window-aligned evaluation become substantially larger when a single alignment is retained across the sequence. Relative to HOI-TG, METAV3RSE reduces CD_h, CD_o, and CD_c by 18.9, 18.7, and 17.7 cm, respectively, with similar improvements over CONTHO. METAV3RSE also obtains the lowest $\mathrm{V2V}_o$ at 43.5 cm. Thus, the improvement in human-object reconstruction is retained when errors accumulated after the initial alignment remain visible to the evaluator.

This result is consistent with the recurrent formulation of METAV3RSE. In Eq. (5), the human, object, contact, and scene representations are decoded jointly with the recurrent state $\mathbf{S}_{t-1}$, providing a shared context across frames. Under the first-frame protocol, the evaluator cannot subsequently re-align the reconstruction to compensate for changes in this coordinate frame. The lower errors in Table 7 therefore provide additional evidence that the reconstructed human-object configuration remains more consistent over the sequence under a fixed initial alignment.

Baseline rescoring. All values in Table 7 are obtained by rescoring predictions from the released HOI-TG and CONTHO models using the same frames and implementation as METAV3RSE. This is important because our rescoring of the released HOI-TG predictions gives 62.0 cm CD_o, whereas 24.7 cm is reported in (Zhang et al., 2026) under a stated first-frame alignment setting. We therefore do not combine externally reported numbers in (Zhang et al., 2026) with our evaluation; all comparisons in Table 7 use predictions scored through the same pipeline.

D BACKBONE SCALING: LIGHTWEIGHT VIT-S/672 VARIANT

The main model uses the ViT-L/896 Human3R configuration. Following the backbone-scaling analysis of UniCon3R (Sur et al., 2026), we additionally instantiate METAV3RSE on the lightweight ViT-S/672 configuration, which reduces both the image resolution and the Multi-HMR encoder size. The proposed architecture is otherwise unchanged: both configurations use the same interaction-aware object initialization and mesh-indexed object pose readout. This analysis therefore examines whether the complete METAV3RSE pipeline can also be instantiated with a substantially smaller Human3R backbone.

Following the evaluation order of the main paper, we first examine global human motion and local human mesh reconstruction, followed by human–object reconstruction. For the human-only comparisons, we additionally include the backbone-matched Human3R and UniCon3R results from the UniCon3R supplementary analysis.

Global human motion. Table 8 compares global human motion across backbone scales. On the held-out EMDB-2 benchmark, the ViT-S/672 variant of METAV3RSE obtains 110.9 mm WA-MPJPE, 293.8 mm W-MPJPE, 2.5% RTE, and 45.3 mm PA-MPJPE, compared with 130.7, 331.6, 2.5%, and 52.9 for the backbone-matched UniCon3R model. The ViT-L/896 configuration further reduces these errors to 87.0, 220.6, 2.2%, and 34.5, respectively, matching the main-paper results.

On RICH, the ViT-S/672 variant obtains 95.1 mm WA-MPJPE, 148.5 mm W-MPJPE, 2.0% RTE, and 51.5 mm PA-MPJPE, compared with 97.2, 151.1, 2.0%, and 51.7 for backbone-matched UniCon3R. The corresponding ViT-L/896 results are 79.1, 126.2, 1.7%, and 37.1.

Local human mesh reconstruction. Table 9 reports local human mesh reconstruction. On 3DPW, the ViT-S/672 variant obtains 93.9 mm PVE, 81.0 mm MPJPE, and 52.5 mm PA-MPJPE, improving

Table 8: **Global human motion estimation across backbone scales on EMDB-2 (Kaufmann et al., 2023) and RICH (Huang et al., 2022a).** We compare the lightweight ViT-S/672 variant with the ViT-L/896 configuration used in the main paper. Human3R* denotes the released checkpoint at the indicated backbone, evaluated in the UniCon3R pipeline, while Human3R† denotes the same architecture and backbone trained with the UniCon3R training protocol (Sur et al., 2026). WA-MPJPE, W-MPJPE, and PA-MPJPE are reported in mm and RTE in %. Dashes denote unavailable measurements. Best results are in **bold** and second-best results are underlined within each backbone group.

Method	EMDB-2 (Kaufmann et al., 2023)				RICH (Huang et al., 2022a)			
	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓	PA-MPJPE ↓
ViT-S/672								
Human3R* (Chen et al., 2026)	137.8	333.1	**2.4**	52.9	131.5	207.3	3.3	71.8
Human3R† (Chen et al., 2026)	–	–	–	–	113.9	178.3	2.6	58.3
UniCon3R (Sur et al., 2026)	130.7	331.6	2.5	52.9	97.2	151.1	**2.0**	51.7
METAV3RSE (Ours)	**110.9**	**293.8**	2.5	**45.3**	**95.1**	**148.5**	**2.0**	**51.5**
ViT-L/896								
Human3R* (Chen et al., 2026)	118.2	286.1	2.4	38.7	109.9	184.0	3.3	48.6
Human3R† (Chen et al., 2026)	–	–	–	–	97.5	153.2	2.3	41.4
UniCon3R (Sur et al., 2026)	113.7	285.6	2.3	38.2	81.5	129.8	**1.7**	37.6
METAV3RSE (Ours)	**87.0**	**220.6**	**2.2**	**34.5**	**79.1**	**126.2**	**1.7**	**37.1**

Table 9: **Local human mesh reconstruction across backbone scales on 3DPW (von Marcard et al., 2018) and EMDB-1 (Kaufmann et al., 2023).** We compare the lightweight ViT-S/672 variant with the ViT-L/896 configuration used in the main paper. All errors are reported in mm. Human3R* denotes the released checkpoint. Best results are in **bold** and second-best results are underlined within each backbone group.

Method	3DPW (von Marcard et al., 2018)			EMDB-1 (Kaufmann et al., 2023)		
	PVE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓
ViT-S/672						
Human3R* (Chen et al., 2026)	103.1	87.8	56.1	107.9	94.7	67.8
UniCon3R (Sur et al., 2026)	103.0	87.5	57.7	119.0	106.3	71.8
METAV3RSE (Ours)	**93.9**	**81.0**	**52.5**	**93.7**	**84.0**	**61.4**
ViT-L/896						
Human3R* (Chen et al., 2026)	86.6	72.8	44.6	86.8	74.6	49.2
UniCon3R (Sur et al., 2026)	86.8	72.5	45.4	94.3	81.7	49.8
METAV3RSE (Ours)	**83.1**	**72.0**	**42.5**	**75.6**	**65.7**	**45.0**

over the backbone-matched UniCon3R model at 103.0, 87.5, and 57.7 mm. With ViT-L/896, the corresponding errors decrease to 83.1, 72.0, and 42.5 mm, consistent with the main paper.

On the held-out EMDB-1 benchmark, the ViT-S/672 model obtains 93.7 mm PVE, 84.0 mm MPJPE, and 61.4 mm PA-MPJPE, while ViT-L/896 obtains 75.6, 65.7, and 45.0 mm, respectively.

Human–object reconstruction. Table 10 compares the two METAV3RSE backbone configurations with the feed-forward baselines used in the main paper. On InterCap, the ViT-S/672 model obtains $\mathrm{CD}_h = 7.2$ cm, $\mathrm{CD}_o = 23.2$ cm, $\mathrm{CD}_c = 8.3$ cm, and $\mathrm{V2V}_o = 31.4$ cm. The corresponding ViT-L/896 results are 7.4, 26.9, 8.7, and 32.9 cm, respectively. Thus, the lightweight configuration is slightly stronger on the InterCap reconstruction metrics.

On BEHAVE, the ViT-L/896 configuration is stronger, obtaining 22.5 cm human CD, 21.6 cm object CD, 18.8 cm combined CD, and 28.4 cm object V2V, compared with 24.5, 22.3, 19.8, and 29.0 cm for ViT-S/672. The corresponding object acceleration errors are 7.3 and 9.4 cm/frame^2, respectively. The relative ordering therefore depends on the benchmark rather than consistently favoring one backbone size.

Together, these results show that the interaction-aware initialization and mesh-indexed object pose readout are not tied to the ViT-L/896 configuration used in the main experiments, and can also be instantiated with the lightweight ViT-S/672 Human3R backbone.

Table 10: **Human–object reconstruction across backbone scales on InterCap (Huang et al., 2022c) and BEHAVE (Bhatnagar et al., 2022).** We compare the lightweight ViT-S/672 and ViT-L/896 configurations of METAV3RSE against the feed-forward baselines used in the main paper. The backbone column applies to METAV3RSE; dashes denote methods that do not use the Human3R backbone family. Reconstruction distances are reported in cm and acceleration errors in cm/frame^2; lower is better. InterCap follows the framewise evaluation protocol and therefore does not report acceleration errors. Best results are in **bold** and second-best results are underlined.

Method	Backbone	**InterCap** (Huang et al., 2022c)				**BEHAVE** (Bhatnagar et al., 2022)					
		$CD_h \downarrow$	$CD_o \downarrow$	$CD_c \downarrow$	$V2V_o \downarrow$	$CD_h \downarrow$	$CD_o \downarrow$	$CD_c \downarrow$	$V2V_o \downarrow$	$Acc_h \downarrow$	$Acc_o \downarrow$
HOI-TG (Wang et al., 2025b)	–	8.7	26.9	9.7	34.5	34.0	22.4	24.6	**25.6**	**1.7**	**5.3**
CONTHO (Nam et al., 2024)	–	8.9	27.8	10.0	35.2	34.3	22.5	24.8	25.7	1.9	5.5
METAV3RSE (Ours)	ViT-S/672	**7.2**	**23.2**	**8.3**	**31.4**	24.5	22.3	19.8	29.0	3.3	9.4
METAV3RSE (Ours)	ViT-L/896	7.4	26.9	8.7	32.9	**22.5**	**21.6**	**18.8**	28.4	2.8	7.3

Table 11: **Human-object and human-scene contact. Left:** Human-object contact on BEHAVE (Bhatnagar et al., 2022) and InterCap (Huang et al., 2022c), evaluated using precision (P) and recall (R). **Right:** Human-scene contact on RICH (Huang et al., 2022a), evaluated using precision (P), recall (R), and F1. Best results are in **bold** and second-best results are underlined.

(a) Human-object contact

Method	BEHAVE		InterCap	
	P ↑	R ↑	P ↑	R ↑
POSA (Hassan et al., 2021)	0.51	0.30	0.56	0.33
BSTRO (Huang et al., 2022a)	0.62	0.53	0.51	0.43
DECO (Tripathi et al., 2023)	0.64	0.34	0.64	0.48
PHOSA (Zhang et al., 2020)	0.39	0.27	0.23	0.16
CHORE (Xie et al., 2022)	0.59	0.47	0.34	0.25
CONTHO (Nam et al., 2024)	**0.75**	0.59	0.66	**0.61**
HOI-TG (Wang et al., 2025b)	0.66	0.55	**0.70**	0.47
ScoreHOI (Li et al., 2025)	0.63	0.59	0.63	0.59
METAV3RSE (Ours)	0.66	**0.64**	0.59	0.43

(b) Human-scene contact

Method	P ↑	R ↑	F1 ↑
DECO (Tripathi et al., 2023)	**0.71**	0.76	0.70
InteractVLM (Dwivedi et al., 2025)	0.63	0.63	0.60
Human3R* (Chen et al., 2026)	0.07	0.13	0.08
Parallel Readout (Sur et al., 2026)	0.59	0.72	0.62
UniCon3R (Sur et al., 2026)	0.64	0.80	**0.71**
METAV3RSE (Ours)	0.64	**0.82**	0.69

E CONTACT EVALUATION

We provide additional evaluation of both human-object and human-scene contact. For human-object contact, we compare against prior HOI and contact-aware methods on BEHAVE (Bhatnagar et al., 2022) and InterCap (Huang et al., 2022c). For human-scene contact, we follow the evaluation protocol of UniCon3R (Sur et al., 2026) on RICH (Huang et al., 2022a). We report precision and recall for human-object contact, and precision, recall, and F1 for human-scene contact.

Human-object contact. On BEHAVE, METAV3RSE achieves the highest recall at 0.64, while matching the second-best precision of 0.66. This indicates that the model recovers a larger fraction of the interacting body region without sacrificing precision relative to most prior methods. On InterCap, METAV3RSE obtains 0.59 precision and 0.43 recall. While dedicated contact-centric methods such as CONTHO (Nam et al., 2024) and HOI-TG (Wang et al., 2025b) achieve higher contact scores on this benchmark, our contact prediction is learned as part of the full human-object-scene reconstruction pipeline rather than as an isolated contact estimator.

Human-scene contact. On RICH, METAV3RSE preserves the strong human-scene contact performance of the UniCon3R backbone. It achieves the highest recall of 0.82, compared with 0.80 for UniCon3R, while maintaining the same precision of 0.64. The resulting F1 score of 0.69 remains close to UniCon3R's 0.71, showing that extending the model with object interaction does not substantially degrade its inherited human-scene contact capability.

Table 12: **Performance under human-relative object motion.** BEHAVE performance stratified by ground-truth human-relative translational speed. Values are equal-weight sequence means and lower is better for all reconstruction metrics.

Motion quartile	Sequences	Rel. speed (cm/s)	$\mathrm{CD}_o \downarrow$ (cm)	$\mathrm{Acc}_o \downarrow$ (cm/frame^2)	$\mathrm{Acc}_h \downarrow$ (cm/frame^2)
Q1	21	18.58	20.14	4.77	2.37
Q2	21	31.87	20.99	6.63	2.79
Q3	21	39.96	22.64	8.43	2.93
Q4	22	61.79	21.86	10.09	3.20

F Failure Analysis

Human-relative object motion. We examine how reconstruction quality changes as the object moves more strongly relative to the human. For each BEHAVE sequence, we compute the median speed of the ground-truth object translation after subtracting the ground-truth human-root translation, and partition the 85 sequences into quartiles according to this human-relative translational speed. Table 12 reports reconstruction error within each quartile.

From the lowest- to the highest-motion quartile, object acceleration error increases by $5.31\,\mathrm{cm/frame}^2$ (95% category-stratified bootstrap interval $[4.65, 6.01]$), whereas human acceleration error increases by only $0.83\,\mathrm{cm/frame}^2$. The difference between these changes is 4.49 with interval $[3.90, 5.13]$. The association remains after controlling for object category, subject, sequence length, and human-root speed, with a standardized coefficient of 0.42 and interval $[0.33, 0.52]$.

By contrast, CD_o changes by only $1.72\,\mathrm{cm}$ from Q1 to Q4, with a bootstrap interval of $[-0.97, 4.39]$. Thus, increasing human-relative object motion is associated much more strongly with temporal object error than with the frame-wise object surface error measured by CD_o. This is consistent with the design of our mesh-indexed object pose readout, which is trained with spatial pose and geometry supervision but no additional temporal objective.

One possible contributor is the temporal coverage of the object-supervised training data. The verified chronological training candidates are highly imbalanced across object categories, with 95.7% belonging to a single category. This motivates future training data that are balanced jointly by object category and human-relative motion. We emphasize that this analysis shows an association and does not establish a causal effect of training-data coverage. We do not extend the analysis to InterCap because its evaluation protocol does not provide the corresponding acceleration metric.

The trend is stable to the aggregation convention: frame-weighted evaluation gives an Acc_o Q4–Q1 difference of 4.98 with interval $[4.39, 5.62]$. The association also remains in leave-one-category-out and leave-one-subject-out adjusted fits.

G Additional Qualitative Results

Global human trajectory estimation on EMDB-2. Figure 4 provides additional qualitative comparisons of global human trajectory estimation on EMDB-2. We visualize the reconstructed human motion in the common world frame together with the corresponding camera trajectory, highlighting differences in global localization and temporal consistency across methods.

Local human mesh recovery. Figure 5 shows additional qualitative comparisons of local human mesh recovery. These examples complement the quantitative results by illustrating differences in body pose and mesh alignment across challenging frames.

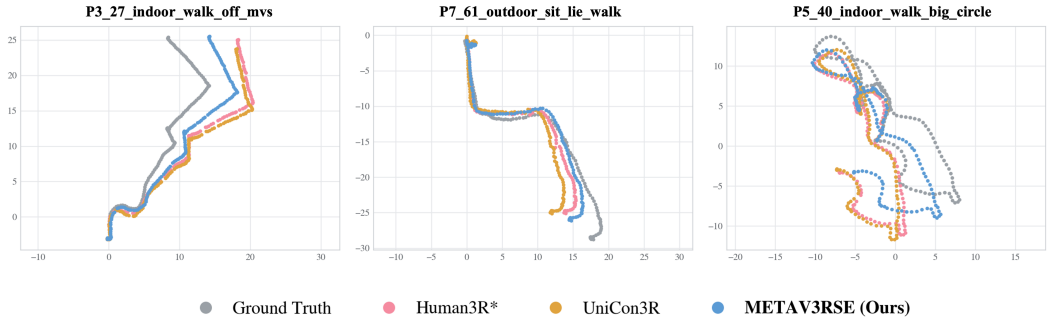


Figure 4: **Comparison of global human trajectory estimation on EMDB-2 (Kaufmann et al., 2023).** We compare Human3R*, UniCon3R and METAV3RSE against ground truth after world-coordinate alignment.



Figure 5: **Qualitative comparison of local human mesh recovery.** We compare Human3R*, UniCon3R and METAV3RSE against ground truth. METAV3RSE produces body pose and scene alignment that are closer to ground truth, particularly in cases with clear scene support.