SCENE AND HUMAN IN ONE WORLD: RECONSTRUCTION IN A FEEDFORWARD PASS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing humans in dynamic scenes from monocular video remains challenging due to scale ambiguity and human–scene misalignment. We observe that a monocular point map is defined only up to a single global scale, which scene geometry alone cannot resolve, whereas the parametric human body has a statistically well-constrained real-world size. We introduce SHOW, a mask-promptable feed-forward framework that reconstructs a designated person together with the surrounding scene in one shared metric space, in which the human body acts as the metric reference that determines the scale of the normalized point map. Mask prompting and auxiliary DensePose supervision make the geometry features discriminative on human regions, and a unified decoder regresses body pose, shape, and global translation together with the scale factor of the scene, so that scale, placement, and scene geometry are optimized as a single system instead of being aligned post hoc. SHOW therefore produces a human reconstruction that is already metrically consistent with the reconstructed scene. On 3DPW, EMDB, and RICH, it improves metric-scale consistency and human–scene alignment over prior joint human–scene feed-forward methods while remaining competitive on global human motion estimation.

1 INTRODUCTION

Humans and scenes coexist in the same physical world: a person's position, size, and contact with the environment are all defined relative to the surrounding geometry. Reconstructing the two in a shared 3D space is therefore essential for applications such as AR/VR, gaming, and robotics (Wang et al., 2023; Yi et al., 2024; Wang et al., 2025a; Xie et al., 2022; 2023; 2026; 2025), where a body that is metrically inconsistent with its environment cannot be used.

Existing pipelines, however, still treat humans and scenes as separate targets that are aligned post hoc (Liu et al., 2026; Xue et al., 2024). They rely on fragile cues such as heuristic body anchors or detector-based estimates, and use scene geometry only as auxiliary context rather than an active constraint for human reconstruction, so the recovered human and scene may look plausible in isolation yet be inconsistent in scale and position when placed together. This is especially severe for monocular human-centric videos, where depth and scale ambiguity makes it difficult to determine whether the body is correctly localized in the surrounding scene. The same separation carries over to evaluation: humans and scenes are scored by separate benchmarks, so strong numbers on both do not imply that the two are consistent.

Recent feed-forward methods reconstruct humans and scenes within a unified pipeline: Human3R (Chen et al., 2025b) recovers humans, scenes, and camera trajectories in a single forward pass using the 3D priors of a visual geometry foundation model, and UniSH (Li et al., 2026) fuses scene and human reconstruction branches for metric-scale reconstruction. These methods, however, operate in a single pipeline rather than in a shared metric space: Human3R decodes body parameters from a geometry backbone trained with additional metric point-map data, and UniSH freezes both branches and estimates scale with a separate alignment network trained on extra in-the-wild video. In both cases, the human model is a SMPL-X mesh (Loper et al., 2015; Pavlakos et al., 2019) regressed by an HMR network (Kanazawa et al., 2018; Goel et al., 2023), which supplies a scale cue that is aligned afterwards, rather than determining the scale of the scene representation itself, so the

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



Figure 1: Our method can produce better human and scene alignment under monocular human-centric video input. Our method will predict human parametric model and scene point map in a feedforward manner.

metric size of the body and the metric scale of the scene are never tied to a single jointly optimized variable.

In contrast, we argue that humans and scenes should be reconstructed in a single metric space. Our key observation is that the scale of a monocular point map is a single unknown scalar, whereas the human body is an object of known metric size; requiring the body to lie on the reconstructed geometry therefore determines that scalar, turning human-scene alignment into a well-posed calibration problem that the human body itself resolves. We therefore propose SHOW, a mask-promptable feed-forward framework for human-scene aligned reconstruction from monocular human-centric videos (Fig. 1). Solving this calibration requires (i) geometry features that identify which part of the reconstruction belongs to the human, and (ii) an explicit, jointly optimized scale variable defined in the camera space shared by both branches.

For (i), SHOW fine-tunes a pretrained visual geometry backbone (Wang et al., 2025b) with a mask-prompting mechanism and an auxiliary DensePose head (Güler et al., 2018; Neverova et al., 2020) so that its geometry features are discriminative on human regions. Human masks serve as dense, instance-level prompts (Wang et al., 2025e; Yang et al., 2026) that identify the target person; unlike prior works that use such prompts only inside the human branch, we feed them into the geometry backbone, so that the scene representation itself is conditioned on the target human. DensePose supervision then injects body semantics into these features: since the same features must predict both the point map and dense body-surface correspondences, they learn to distinguish human-region pixels from background. This does not make the predicted geometry more accurate, but it makes the human region identifiable within it, so that scale estimation operates on features specific to the body rather than to the surrounding scene.

For (ii), SHOW predicts an explicit scale variable that ties the two. Since the body and the point map can only be compared if they live in the same camera space, we condition the human branch on the camera parameters estimated by the geometry backbone through ray-based positional encodings, and fuse HMR features with the human-aware geometry features. A lightweight decoder then predicts body pose, shape, and translation together with a scale factor for the scene representation, which lifts the normalized point map to metric scale; joint training supervises this scale directly while pulling the rescaled human-region point map onto the reconstructed body surface, so that scale, placement, and scene geometry are optimized as a single system.

Our contributions are summarized as follows:

- We introduce SHOW, a unified feed-forward framework that jointly estimates dense scene geometry and the mesh of a designated person in a shared metric coordinate system, reducing scale ambiguity and the cascading errors of multi-stage alignment pipelines.
- We cast human-scene reconstruction as a joint scale-calibration problem: the body provides a metric reference that determines the scale of the normalized point map, so that the body, its global placement, and the scene geometry are tied to a single jointly optimized variable.

- We propose a mask-as-prompt mechanism that uses human masks as dense 2D prompts for visual geometry foundation models, so that the scene representation itself is conditioned on the target person rather than only the human branch.
- We provide a comprehensive evaluation protocol for human-scene alignment, with extensive benchmarking of existing feed-forward methods.

2 Related Work

2.1 Human Mesh Recovery

Human mesh recovery (HMR) estimates body pose and shape in a parametric body model. SMPL (Loper et al., 2015) and SMPL-X (Pavlakos et al., 2019) provide a metrically scaled mesh template with a learned low-dimensional prior over plausible poses, which makes single-image body estimation well-posed and gives the recovered body an absolute metric size. Camera-space methods regress these parameters from a cropped image (Kanazawa et al., 2018; 2019; Kocabas et al., 2020), and recent variants improve accuracy with tokenized pose representations (Dwivedi et al., 2024), transformer-based whole-body regression (Goel et al., 2023; Yang et al., 2026), training that aligns the body with the estimated perspective (Patel & Black, 2025), single-shot multi-person regression (Baradel et al., 2024), and prompt-based identification of a target person in crowded scenes (Wang et al., 2025e). These methods recover local pose and shape well, but they operate in the camera frame, leaving the metric scale, depth, and global placement of the body unresolved. World-grounded HMR lifts motion into a gravity-aware world frame (Shin et al., 2024; Yuan et al., 2022; Wang et al., 2024b), typically through relative motion prediction or global camera rotation estimation (Shen et al., 2024; Ye et al., 2023), but it does not align the body with the surrounding scene geometry.

2.2 Feed-Forward 3D Reconstruction

DUSt3R (Wang et al., 2024a) regresses pixel-aligned point maps from an image pair, and its successors extend the paradigm to multi-view and streaming inputs with explicit spatial memory (Wang & Agapito, 2025; Wu et al., 2025; Zhuo et al., 2026), to dynamic scenes without test-time optimization (Zhang et al., 2025; Chen et al., 2025a), and to long sequences through test-time training (Chen et al., 2026). VGGT (Wang et al., 2025b) predicts camera parameters, depth, and point maps jointly in a single transformer, π^3 (Wang et al., 2025d) removes the fixed reference view, and Depth Anything 3 (Lin et al., 2025) recovers visual space from arbitrary views. Because these models encode strong camera and geometry priors, recent work reuses them for human reconstruction: Human3R (Chen et al., 2025b) decodes SMPL-X bodies from the latent state of the online reconstruction model CUT3R (Wang et al., 2025c), while HAMSt3R (Rojas et al., 2025) augments MASt3R (Leroy et al., 2024) with instance and DensePose heads to obtain human-aware point maps. In these systems, however, the human prior enters only at the decoder or through an auxiliary head, and associating a person with the reconstructed scene still relies on coarse cues such as head detection or bounding boxes, which are unreliable under occlusion, truncation, and crowded scenes.

2.3 Joint Human-Scene Reconstruction

Optimization-based methods reconstruct the scene first and then fit or optimize the human within it. Early works either populate reconstructed scenes with plausible human bodies or synthesize scenes conditioned on human poses (Hassan et al., 2021a;b; Cao et al., 2026; Wang et al., 2025a; Li et al., 2024; Nie et al., 2021). More recent methods jointly optimize humans, cameras, and scene geometry: JOSH (Liu et al., 2026) refines camera poses, global human motion, and a dense point cloud in a single stage, using human-scene contact constraints to anchor metric scale, while Müller et al. (2025) place humans inside a structure-from-motion pipeline and show that human data also improves camera pose and scene reconstruction. These methods align humans and scenes accurately but require per-scene optimization, depend on initialization from off-the-shelf models, and cannot run online.

Feed-forward methods predict humans and scenes in one pass. Human3R (Chen et al., 2025b) and JOSH3R (Liu et al., 2026) recover global multi-person meshes together with dense scene geometry,

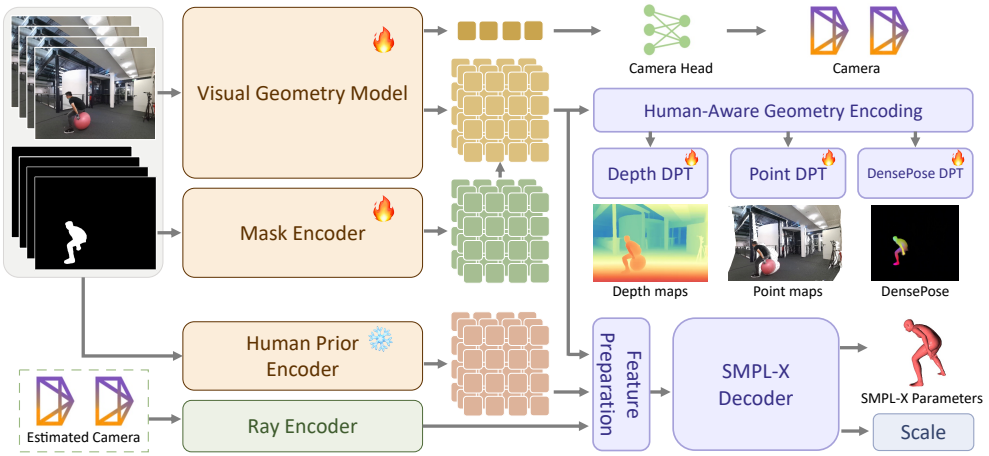


Figure 2: Given a monocular video and a mask prompt, SHOW reconstructs the target person and the surrounding scene in a shared 3D world in one feed-forward pass. VGGT is adapted with mask prompting and an auxiliary DensePose head to extract human-aware geometry features and predict normalized scene point maps. A geometry-aware SMPL-X decoder then estimates body pose, shape, global translation, and the scale factor that lifts the normalized point map to metric scale.

HAMSt3R (Rojas et al., 2025) reconstructs humans and scenes from sparse multi-view images, VideoMimic (Allshire et al., 2025) reconstructs scene-aware human motion for humanoid control, and UniSH (Li et al., 2026) keeps both branches frozen and predicts scale and translation with a lightweight alignment network. They are fast and temporally coherent, but the metric scale of the scene and the metric size of the body come from separate components, so their consistency is only approximate—the coupling that we target.

3 METHOD

Given a monocular video $\mathcal{I} = \{I_t\}_{t=1}^{T}$ and a mask prompt m_t that identifies the target person in each frame, our goal is to reconstruct that person and the surrounding scene together in a shared 3D world in a single feed-forward pass. For each frame I_t, we predict the scene representation $\mathcal{S}_t = (\bar{\mathbf{P}}_t, s_t)$, consisting of a normalized partial 3D point map $\bar{\mathbf{P}}_t = \{\bar{\mathbf{p}}_{t,u}\}_{u\in\Omega}$ and a scalar s_t that maps it to metric scale, $\mathbf{P}_t = s_t\bar{\mathbf{P}}_t$, together with the prompted human $H_t = \{\theta_t, \beta_t, \tau_t\}$ in the same coordinate frame. Here Ω denotes the set of image pixels and $\bar{\mathbf{p}}_{t,u} \in \mathbb{R}^3$ is the normalized 3D point corresponding to pixel u; θ_t, β_t, and τ_t denote pose, shape, and global translation, respectively. Estimating s_t therefore places the human and the scene in a single metric coordinate system, and allows the reconstructed human to be geometrically aligned with the surrounding scene.

Figure 2 provides an overview of our pipeline. First, we adapt VGGT (Wang et al., 2025b) for human-aware geometry encoding by incorporating a mask encoder and an auxiliary DensePose head (Sec. 3.1). Second, we train a geometry-aware SMPL-X decoder that combines human features (Wang et al., 2025e) with scene context to recover body parameters in the reconstructed 3D space (Sec. 3.2). Finally, we jointly fine-tune the human and scene reconstruction modules so that the recovered body's scale and placement are aligned with the surrounding 3D scene (Sec. 3.3).

3.1 HUMAN-AWARE GEOMETRY ENCODING

VGGT (Wang et al., 2025b) provides strong feed-forward geometry features for camera, depth, and point-map prediction, but its latent features are optimized for generic scene geometry and carry no notion of which region of the image is the target person. Since the prompted human occupies only a small fraction of the point map, its depth is not emphasized in these features, even though that depth is exactly what must be aligned with the placement of the reconstructed body. We therefore adapt the geometry encoding so that its latent features retain human-related cues for downstream SMPL-X decoding: a lightweight mask encoder is incorporated, and the latent features are supervised with an auxiliary DensePose head so that they learn to recognize human body regions in the reconstructed 3D space.

Mask Encoder. For each frame I_t, the mask prompt is a human foreground mask m_t, obtained with an off-the-shelf segmentation model (Ravi et al., 2024; Carion et al., 2025). We encode it with a lightweight mask encoder $F_t^m = E_{msk}(m_t)$. In parallel, the VGGT image encoder E_{geo} extracts latent geometry features $F_t^g = E_{geo}(I_t)$ from the same frame. We incorporate the mask feature into the latent geometry feature through an adaptation module, $\tilde{F}_t^g = A(F_t^g, F_t^m)$, where $\tilde{F}_t^g$ is the resulting human-aware geometry feature. The adapted feature $\tilde{F}_t^g$ is then used for depth and point-map prediction, and later passed to the SMPL-X decoder so that the depth of the human region informs the estimated scale and placement of the body.

DensePose Auxiliary Supervision. A mask encoder alone does not guarantee that the geometry features become human-aware, since the network may still treat the mask as a generic foreground cue. We therefore supervise the latent feature $\tilde{F}_t^g$ with an auxiliary DensePose prediction task, using a regression loss that upweights human pixels so that it is not dominated by the background.[1] This forces the shared feature to preserve information predictive of human body regions rather than merely separating foreground from background. We verify by ablation that part-level body semantics, rather than a generic foreground cue, is what makes the geometry features discriminative around the human (Sec. 4.4). The head is used only during training and discarded afterwards. This differs from HAMSt3R (Rojas et al., 2025), which keeps a DensePose head at inference time: our purpose is to make the geometry features discriminative on the human region, so that the predicted scale is grounded in body-specific rather than scene-generic evidence.

We train the human-aware geometry encoding with both the original VGGT geometry losses and the auxiliary DensePose loss, $\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pmap}} + \lambda_{\text{dp}}\mathcal{L}_{\text{dp}}$: the DensePose loss injects human-specific supervision, while the geometry losses, which follow VGGT (Wang et al., 2025b), prevent degradation in camera, depth, and point-map prediction.[2] The resulting features retain VGGT's scene-level geometry capability while becoming more useful for downstream SMPL-X decoding.

3.2 Geometry-Aware SMPL-X Decoding

After obtaining the human-aware geometry features from Sec. 3.1, we decode SMPL-X parameters with explicit scene context. As shown in Fig. 3, the decoder takes two complementary feature streams: human-centric features that provide body pose and shape cues, and adapted geometry features that provide scene context for scene scale and human placement.

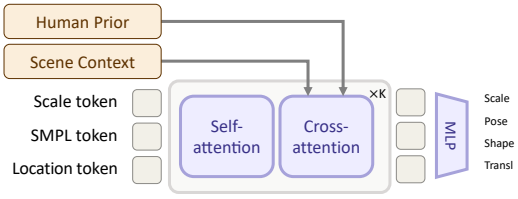


Figure 3: SMPL-X decoder architecture. Scale, SMPL, and location tokens pass through self- and cross-attention with geometry and human patch features, and are decoded into scale, pose, shape, and translation.

Feature Preparation. For frame I_t, we extract human feature tokens with the frozen DINOv2 backbone (Oquab et al., 2023) of PromptHMR (Wang et al., 2025e):

$$F_t^h = E_{\text{phmr}}(I_t, m_t), \tag{1}$$

where m_t is the mask prompt. We pair these tokens with the adapted geometry feature $\tilde{F}_t^g$ from Sec. 3.1 and add ray-based positional encodings computed from the camera intrinsics $\hat{\mathbf{K}}_t$ predicted by the camera head of the scene branch, $\bar{F}_t^h = F_t^h + \text{RayEnc}(\hat{\mathbf{K}}_t)$, so that the human tokens are expressed in the same camera frame as the scene. The resulting human tokens $\bar{F}_t^h$ and geometry tokens $\tilde{F}_t^g$ are used as inputs to the geometry-aware SMPL-X decoder.

Geometry-Aware Token Decoder. We use a lightweight Transformer decoder to regress SMPL-X parameters from the prepared human and geometry tokens. We initialize three query tokens $Q_t^0 = \{q_t^{\text{smpl}}, q_t^{\text{scale}}, q_t^{\text{loc}}\}$: the SMPL token q_t^{smpl} is responsible for body pose and shape, the scale token q_t^{scale} estimates the point-map scale parameter, and the location token q_t^{loc} predicts the body's location in the camera space of frame I_t. Each decoder layer first performs self-attention

[1] The exact form is given in Eq. 5 of Sec. C in the supplementary material.

[2] All terms are defined in Sec. C of the supplementary material.

among the three query tokens and then cross-attends to the human and geometry features, $Q_t^{l+1} = \mathrm{CrossAttn}(\mathrm{SelfAttn}(Q_t^l); \bar{F}_t^h, \tilde{F}_t^g)$, so that every token draws on both the human prior and the scene context. We use $K = 4$ layers, each cross-attending to the geometry features of a different VGGT layer (4, 11, 17 and 23); allowing the human branch to interact with scene knowledge at several levels of abstraction is preferable to attending to the last layer alone (Sec. 4.4).

After K decoder layers, three lightweight MLP heads regress the final parameters from their corresponding tokens: body pose and shape ($\hat{\theta}_t, \hat{\beta}_t$), global translation ($\hat{\tau}_t$), and the scale of the scene representation ($\hat{s}_t$), which lifts the normalized point map to metric scale.

Decoder Training. In the second training stage, we freeze the human-aware geometry encoder and train only the SMPL-X decoder with standard human reconstruction losses $\mathcal{L}_{\mathrm{hmr}}$, a weighted sum of 2D and 3D keypoint losses, SMPL-X parameter losses, and vertex and translation losses.[3] This stage teaches the decoder to recover body pose, shape, translation, and scale from the combined human and geometry features, while the next stage further enforces their alignment with the reconstructed scene.

3.3 Joint Training for Human-Scene Alignment

After the first two stages, the geometry encoder provides human-aware scene features and the SMPL-X decoder predicts the human body from both human and scene cues, but training the two components separately still does not explicitly enforce consistency between the reconstructed body and the predicted 3D point map. We therefore perform a final joint fine-tuning stage to align the human reconstruction with the surrounding scene geometry.

Scale and Placement Consistency. The predicted scale $\hat{s}_t$ rescales the normalized point map so that it is consistent with the reconstructed body, and the predicted translation $\hat{\tau}_t$ places the body in the camera space of frame I_t. We supervise the scale in log space against the ground-truth scale s_t^* and, since a sequence depicts a single scene whose metric scale is constant, regularize it toward its sequence-level average.

SMPL-Guided Point-Map Alignment. We also use the reconstructed body surface to supervise the scale estimate. We render the predicted SMPL-X mesh to obtain its visible surface points $\hat{V}_t^{\mathrm{vis}}$, rescale the human-region point map $\hat{\tilde{\mathbf{P}}}_t[m_t]$ by the predicted scale, and align the two with a pixel-level chamfer loss. Since the visibility operation is non-differentiable, the visible points are detached and serve as a fixed target, so that the gradient updates the predicted scale and the scene branch's point map jointly and the two branches converge on a scale at which the human and the scene point map coincide.

Joint Fine-tuning Objective. The final objective combines the geometry loss, the SMPL-X reconstruction loss, and the human-scene alignment losses, $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{hmr}} + \lambda_{\mathrm{scale}}\mathcal{L}_{\mathrm{scale}} + \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}$.[4] This stage jointly fine-tunes both modules, encouraging the normalized point map, the predicted scale factor, and the reconstructed body surface to be consistent in the shared 3D space.

4 Experiments

4.1 Implementation Details

We train the visual geometry backbone, then the SMPL-X decoder, and finally the full model end-to-end, using 4, 2 and 4 NVIDIA B200 GPUs for 100, 100 and 200 epochs respectively (Secs. 3.1–3.3).

4.2 Datasets and Metrics

Datasets. We train our model on **BEDLAM2.0** (Tesch et al., 2025), a large-scale synthetic dataset that provides ground truth SMPL-X parameters, color images, depth annotations, and camera trajectories. To ensure reliable geometric supervision, we discard sequences with inaccurate HDRI-based

[3] These losses are defined in Sec. C of the supplementary material.

[4] The individual terms are defined in Sec. D of the supplementary material.

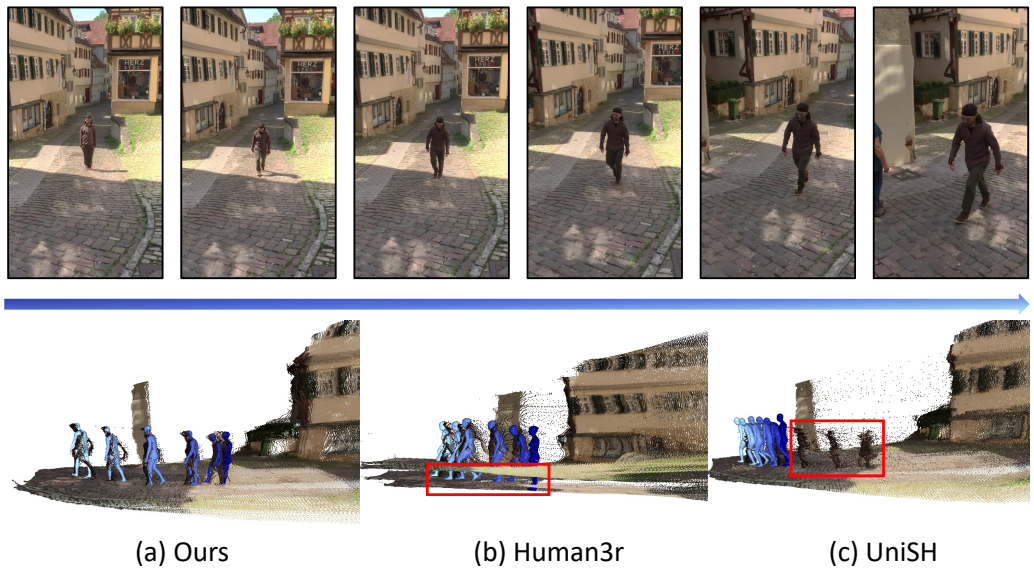


Figure 4: Qualitative comparison of human point map and body model alignment. Given a monocular human-centric video (top), we compare our method with Human3R and UniSH. Our method aligns accurately with the human point map, whereas Human3R shows noticeable ground penetration of the body model and UniSH fails to align over large regions.

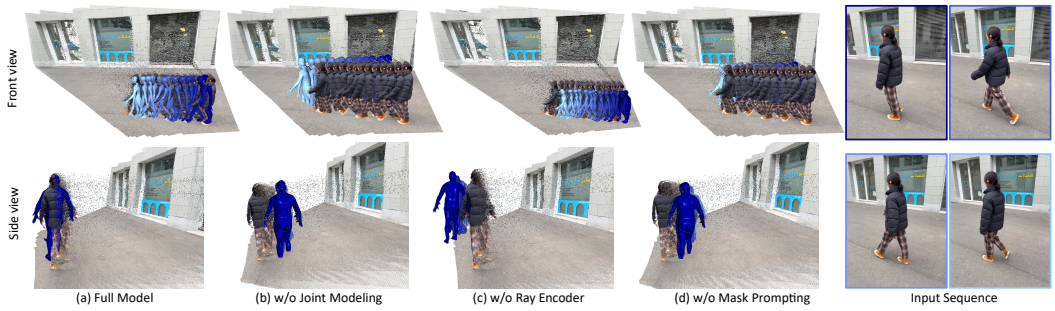


Figure 5: Ablation study of our design choices. (a) Full model prediction, where the human is well aligned with the scene point cloud. (b) Estimating scale solely from scene semantics without the human branch fails to place human and scene in a consistent global space. (c) Without the ray encoder, the human branch cannot incorporate scene information or recover consistent camera intrinsics. (d) Without mask prompting, scale prediction is affected by irrelevant regions.

depth rendering, resulting in 4,250 valid sequences, and downsample videos to 6 FPS with 5-frame clips. For fairness, we note that the baselines in Table 3 use additional data—Human3R trains its scene branch on metric-scale point-map data and UniSH trains its scale branch on self-collected real-world video—whereas SHOW is trained only on the open-source BEDLAM2.0, which makes our comparison conservative.

Metrics. For human mesh recovery, we follow standard evaluation protocols in both camera and world coordinate spaces. In camera space, we evaluate local pose and shape accuracy on 3DPW (von Marcard et al., 2018) and EMDB (subset 1; Kaufmann et al., 2023) with MPJPE, PA-MPJPE, and PVE (mm). In world space, we evaluate long-term motion on EMDB (subset 2) and RICH (Huang et al., 2022), splitting each sequence into 100-frame segments and reporting World MPJPE (W-MPJPE), World-aligned MPJPE (WA-MPJPE), and Root Translation Error (RTE, %), which assess global trajectory accuracy.

To evaluate human–scene consistency, we propose two new metrics: **Human–Scene Chamfer (HS-CF)**, the normalized chamfer distance between body model points and the human point map, and **Human–Scene Variance (HS-V)**, which compares their spatial variance along the horizontal and vertical axes in camera frame (subscripts denote the percentile range used to filter mask-boundary

Cat.	Method	3DPW			EMDB-1		
		PA	MPJ	PVE	PA	MPJ	PVE
w/o sc.	CLIFF (Li et al., 2022)	43.0	69.0	81.2	68.3	103.3	123.7
	HMR2.0a (Goel et al., 2023)	44.4	69.8	82.2	61.5	97.8	120.0
	TokenHMR (Dwivedi et al., 2024)	44.3	71.0	84.6	55.6	91.7	109.4
	CameraHMR (Patel & Black, 2025)	38.5	62.1	72.9	43.7	73.0	85.4
	NLF (Sárándi & Pons-Moll, 2024)	37.3	60.3	71.4	41.2	**69.6**	**82.4**
	PromptHMR (Wang et al., 2025e)	**36.6**	**58.7**	**69.4**	**41.0**	71.7	84.5
w/ sc.	Human3R (Chen et al., 2025b)	44.1	71.2	84.9	48.5	**73.9**	**86.0**
	UniSH (Li et al., 2026)	48.8	75.6	88.8	86.6	112.9	140.0
	Ours	**41.0**	**67.7**	**78.7**	**45.8**	75.7	88.8

Table 1: **Local mesh reconstruction results.** PA: PA-MPJPE (mm). Lower is better for all metrics.

Method	Opt.	Free	Scene	EMDB-2			RICH		
				WA↓	W↓	RTE↓	WA↓	W↓	RTE↓
SLAHMR (Ye et al., 2023)	✗	✗		326.9	776.1	10.2	132.2	237.1	6.4
TRAM (Wang et al., 2024b)	✗	✗		76.4	222.4	1.4	127.8	238.0	6.0
JOSH (Liu et al., 2026)	✗	✓		**68.9**	**174.7**	**1.3**	**89.0**	**132.5**	**3.0**
TRACE (Sun et al., 2023)	✓	✗		429.0	1702.3	17.7	238.1	925.4	101.4
WHAM (Shin et al., 2024)	✓	✗		135.6	334.8	6.0	108.4	190.1	4.5
GVHMR (Shen et al., 2024)	✓	✗		**111.0**	**276.5**	**2.0**	**78.8**	**126.3**	**2.4**
JOSH3R (Liu et al., 2026)	✓	✓		220.0	661.7	13.1	–	–	–
Human3R (Chen et al., 2025b)	✓	✓		112.2	267.9	2.2	110.0	184.9	3.3
UniSH (Li et al., 2026)	✓	✓		118.5	270.1	5.8	118.1	183.2	4.8
Ours	✓	✓		**109.1**	**262.3**	**2.1**	**107.3**	**172.7**	**2.2**

Table 2: **Global human motion estimation.** Opt. Free: feed-forward (optimization-free); Scene: joint 3D scene reconstruction; WA: World-aligned MPJPE; W: MPJPE.

outliers).[5] The two are complementary: HS-CF is normalized by the metric extent of the human and is therefore invariant to a global rescaling of the human–scene pair, whereas HS-V is expressed in absolute metric units and thus also exposes errors in the estimated metric scale itself—the quantity our method targets. We therefore report both metrics, and confirm absolute metric correctness separately with world-space MPJPE (Table 2).

4.3 Comparison

Local human mesh recovery. We compare our method with state-of-the-art human mesh recovery approaches, as shown in Table 1. Our method achieves the best performance on most metrics among joint human–scene feed-forward approaches, because Human3R relies on head detection, which fails under occlusion or uncommon poses, and UniSH uses bounding-box prompts, which are ambiguous when people overlap, whereas our full-body masks specify the target person directly.

Global human mesh recovery. We compare our method with prior approaches for global human mesh recovery in Table 2. Our method outperforms existing joint human–scene feed-forward approaches and matches state of the art; on EMDB-2, which features moving cameras, modeling human and scene in a unified space gives more accurate global motion estimation than methods such as GVHMR that rely on an off-the-shelf scale estimator. On RICH our performance is slightly lower, which we attribute to its static-camera setup: visual geometry models benefit from multi-view observation, which RICH does not provide.

Human scene alignment. We compare our method with recent joint human–scene feed-forward approaches in Table 3. UniSH freezes both branches and estimates scale with a lightweight alignment network, whereas we fine-tune the geometry backbone so that its depth distribution matches the training data and the two branches are more tightly coupled, which yields better human–scene consistency even without additional real-world video. Human3R employs a metric geometry backbone that gives strong HS-CF, but it injects human tokens only at the decoder stage and never fully enforces consistency between the two representations; this is exactly the gap that the two metrics separate. Human3R obtains a more favorable HS-CF than ours (1.30 vs. 5.69), yet places the human at a clearly wrong metric scale, with an HS-V more than twenty times larger (0.652 vs. 0.030). The

[5] Formal definitions are given in Sec. A of the supplementary material.

Method	3DPW				EMDB-1			
	HS-V_5	HS-V_{10}	HS-CF_5	HS-CF_{10}	HS-V_5	HS-V_{10}	HS-CF_5	HS-CF_{10}
Human3R	0.652	0.640	1.30	1.32	3.66	2.91	1.13	1.31
UniSH	**0.028**	0.028	7.65	7.52	0.033	0.037	7.10	6.99
Ours	0.030	**0.026**	**5.69**	**5.44**	**0.013**	**0.017**	**5.88**	**5.79**

Table 3: Comparison of human-scene consistency metrics on 3DPW and EMDB1.

Method	3DPW				
	HS-V_5 ↓	HS-V_{10} ↓	HS-CF_5 ↓	HS-CF_{10} ↓	RTE↓
w/o pretrain alignment	0.708	0.660	38.51	38.39	5.2
w/o joint modeling	0.312	0.295	17.32	17.08	3.3
w/o mask prompting	0.033	0.035	14.12	14.10	2.4
w/o ray encoder	0.734	0.667	11.23	11.06	2.9
full	**0.028**	0.030	9.44	9.38	2.3
full + e2e tuning	0.029	**0.027**	**6.21**	**5.99**	**1.9**

Table 4: Human–scene consistency metrics on 3DPW. Ablation models are trained for 150 epochs on subsampled 6 FPS data; end-to-end fine-tuning adds 100 epochs.

qualitative comparison in Fig. 4 shows the same failure mode, and further examples are given in Fig. 2 of the supplementary material. Our ray encoder instead conditions human prediction on scene intrinsics, which is critical for coherent reconstruction.

4.4 Ablation Study

Table 4 ablates our design choices on 3DPW; Sec. F of the supplementary material analyzes each variant. Removing pretraining alignment degrades every metric severely, since the pretrained backbone uses a different scale convention from the normalized one. Removing joint modeling, the mask prompt, or the ray encoder causes progressively smaller but still clear degradations, confirming that scene-only cues cannot resolve scale ambiguity and that both mask prompting and camera-aware human features are necessary. End-to-end tuning performs best overall; visualizations are shown in Fig. 5. Table 2 of the supplementary material ablates two further design choices and confirms the same picture: replacing the DensePose target with a binary human/background map and attending to the last VGGT layer only both fall clearly below the full model, showing that part-level body semantics and multi-level scene features each contribute to the alignment.

5 Conclusions

We presented **SHOW**, a unified feed-forward framework for joint human–scene reconstruction in a shared metric space from monocular video. Unlike prior methods that treat the two tasks separately, SHOW integrates them within a single model: a mask-as-prompt mechanism guides a pretrained geometry foundation model toward human-relevant 3D structures, and a unified decoder jointly models scene geometry and SMPL-X parameters, so that geometry, camera pose, and human motion are reasoned about consistently. Experiments on 3DPW, EMDB show that SHOW improves metric-scale consistency and human–scene alignment over prior joint human–scene feed-forward methods, while remaining competitive on global human motion estimation.

6 Limitation and Future Work

While our method reconstructs a person and the surrounding scene within a unified metric space, a sim-to-real gap persists: despite large-scale synthetic training data, the model still struggles on out-of-distribution real-world video, such as extreme scale variation or frames in which the human occupies only a small portion of the image. Extending the framework to these regimes, and to finer-grained physical interaction such as contact and penetration, is an important direction for future work.

Reproducibility Statement

We describe the datasets, training stages, and evaluation protocol in Sec. 4.2. All datasets we use are publicly available: we train on BEDLAM2.0 and evaluate on 3DPW, EMDB, and RICH. Formal definitions of the human–scene consistency metrics are given in Sec. A of the supplementary material, the reconstruction and joint-training losses in Secs. C and D, and the architecture and training schedule in Secs. K and L. That document also reports per-component efficiency measurements in Table 1 and a per-variant ablation analysis in Sec. F.

AI Use Statement

In this work, we used generative AI tools for grammar checking, LaTeX formatting, and format checks, and for assistance in editing and revising the manuscript text. The AI-assisted manuscript was inspected for formatting and layout issues. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2025.

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 (ECCV)*, 2024.

Yukang Cao, Haozhe Xie, Fangzhou Hong, Long Zhuo, Zhaoxi Chen, Liang Pan, and Ziwei Liu. HSImul3R: Physics-in-the-loop reconstruction of simulation-ready Human–Scene interactions. *arXiv*, 2603.15612, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *International Conference on Computer Vision (ICCV)*, 2025a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *International Conference on Learning Representations (ICLR)*, 2026.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025b.

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

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, October 2021a.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021b.

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 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

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.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024.

Jianan Li, Tao Huang, Qingxu Zhu, and Tien-Tsin Wong. Physics-based scene layout generation from human motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–10, 2024.

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, 2026. URL `https://arxiv.org/abs/2601.01222`.

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 *ECCV*, 2022.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

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. URL `https://openreview.net/forum?id=7eLE4mfEpz`.

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

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

Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Pose2Room: Understanding 3D scenes from human activities. *arXiv preprint arXiv:2112.03030*, 2021.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, 2025.

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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Sara Rojas, Matthieu Armando, Bernard Ghanem, Philippe Weinzaepfel, Vincent Leroy, and Grégory Rogez. HAMSt3R: Human-aware multi-view stereo 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

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

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

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, 2025.

Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, Qifeng Chen, Jingbo Wang, and Jiangmiao Pang. PhysHSI: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025a.

Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13031–13040, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025d.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024b.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025e.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (ECCV)*. Springer, October 2022.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. InterTrack: Tracking human object interaction without object templates. In *International Conference on 3D Vision (3DV)*, 2025. URL https://arxiv.org/abs/2408.13953.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinhua 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 (ECCV)*, 2024.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. SAM 3D body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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

Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *ECCV*, 2024.

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 (CVPR)*, 2022.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, 2025.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. In *International Conference on Learning Representations (ICLR)*, 2026.