SHOT3R: DECOUPLED MULTI-SHOT REASONING FOR STREAMING 4D HUMAN-SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of Shot3R.** Shot3R reconstructs people, scene, and camera motion from a multi-shot monocular video while streaming. It keeps them in the shared world frame and preserves spatial and identity consistency across shot changes.

ABSTRACT

Multi-shot video splices shots taken at different viewpoints over a continuous timeline. A **shot change** jumps the viewpoint, breaking identities and spatial reference at the shot change. Reconstructing people and scene from such video requires spatial and identity consistency across the changes. Most existing methods resolve cross-shot ambiguities by optimizing all shots and the scene jointly, once the complete sequence has been reconstructed. Unlike such offline methods, streaming reconstruction can only infer such relations from current and past observations, with no global information and only a history from the previous viewpoint. To address this problem, we present **Shot3R**, a streaming framework for 4D reconstruction of multiple people and scene from monocular video. At each shot change, Shot3R infers the transform between consecutive shots in latent space, from the boundary frame's features and the accumulated history, placing the new shot in the shared world frame. Identities are associated across the shot change. The transform is applied to the whole new shot, freeing it from the influence of the previous shot. On extensive multi-shot data, Shot3R delivers marked improvements in human reconstruction, camera pose estimation, and cross-shot identity consistency, largest under large viewpoint changes.

1 INTRODUCTION

Recovering humans and scenes together helps us understand how people move and interact with their environment in 3D (Hassan et al., 2019; Liu et al., 2026). Recent advances enable the recov-

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.

ery of human pose and shape from images (Kanazawa et al., 2018; Goel et al., 2023) and human motion in world coordinates from monocular videos (Yuan et al., 2022; Shin et al., 2024). These methods, however, assume one continuous take, while much of the video we wish to reconstruct, from films and broadcast sports to edited online content, is assembled from shots taken at different viewpoints (Pavlakos et al., 2022). Each such shot change abruptly alters both what is visible and where the camera stands, and the reconstruction must remain consistent across it. For video containing multiple people, this requires expressing camera poses, human motion, and scene geometry in a shared world frame while preserving person identities across shots.

Offline multi-shot methods resolve cross-shot relations over the whole sequence, so their computational cost grows with the length of the video (Zhang et al., 2025b; Kim et al., 2025; On et al., 2026). Streaming recurrent and causal models instead reconstruct incrementally as frames arrive and need no pass over the whole sequence (Wang et al., 2025b; Chen et al., 2026b; Zhuo et al., 2025), but they are built for continuous camera motion and do not treat the shot change itself. The two classes of method each have limitations, and maintaining cross-shot consistency without the complete sequence remains an open problem.

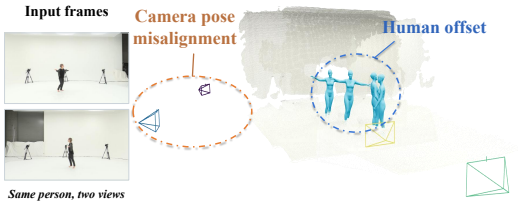

Figure 2: **Spatial inconsistency in multi-shot reconstruction.** Reconstruction with misaligned camera poses and human positions.

A shot change poses two problems, the camera pose of the new view and person identity across the change. As Figure 2 illustrates, propagating recurrent state across an abrupt viewpoint change can produce erroneous camera poses and displaced human reconstructions, because that state comes from the view that just ended. Reconstructing each shot with an independent state avoids inheriting incompatible history, but leaves the shots unrelated, so neither spatial alignment nor identity association is available. TTT3R and Human3R reset the state periodically to limit drift over long sequences (Chen et al., 2026a;b), an adjustment that informs our approach. Their resets, however, are not tied to shot changes, and the camera-based alignment they use does not establish identity associations. Each shot may then be reconstructed correctly in isolation while the sequence remains globally inconsistent.

Building on this observation, our key insight is that historical context can inform cross-shot spatial reasoning without conditioning the new shot's reconstruction. **Shot3R** turns it into three operations. **First**, a cut detector locates the shot change, and the boundary frame is used twice, along a history-conditioned path that estimates the camera pose of the new view in the preceding shot's coordinates and a second path that resets the recurrent state and reconstructs the frame independently. Because both estimates describe the same frame, the transform between the shots follows in closed form, with no correspondence search and no joint optimization. **Second**, relation tokens read the boundary frame's features together with the accumulated history, and the decoded relation corrects the camera latent, placing the new shot in the shared world frame. **Third**, identity is established from body configuration rather than appearance, and the matched pairs yield a shared translation that refines the alignment of cameras, people, and scene points (Figure 1).

Across the datasets we evaluate, Shot3R improves camera pose estimation, human reconstruction, and cross-shot identity consistency. On EgoBody, camera ATE falls from 0.124 m to 0.016 m relative to the continuous-state reference, while IDF1 reaches 0.985; across the EgoHumans strata the advantage grows with the viewpoint change, reaching about 30% lower geometric error in the largest stratum. Controlled comparisons separate the three mechanisms, and the full system runs **5.7×** faster than JOSH at **1.4×** the cost of the continuous-state reference.

Our contributions are summarized as follows:

- We present **Shot3R**, the first framework for streaming 4D human-scene reconstruction from multi-shot video. It handles each shot change as it occurs, placing each new shot in a shared world frame without future frames or sequence-level optimization.
- We propose a **cross-shot alignment mechanism** that separates the two roles history plays at a shot change. It yields the inter-shot transform in closed form without conditioning the new shot's reconstruction, and derives identity from body configuration.

- We report results on EgoBody, EgoHumans, and Harmony4D. Against the continuous-state reference, Shot3R improves human reconstruction and camera pose estimation and preserves identities across shot changes, most of all under large viewpoint changes.

2 Related Work

Human motion and human-scene reconstruction. Human reconstruction has progressed from camera-relative mesh estimation from images (Kanazawa et al., 2018; Kolotouros et al., 2019; Goel et al., 2023) and videos (Kocabas et al., 2020) to motion recovery in a shared world frame. Human and camera motion are coupled in monocular video, so this shift requires separating the movement of the camera from that of the people it observes. Methods in this line therefore pair a learned motion prior with an explicit camera estimate (Yuan et al., 2022; Ye et al., 2023; Shin et al., 2024). Subsequent work reduces the remaining ambiguity through metric-scale camera recovery (Wang et al., 2024b) and gravity-view pose representations (Shen et al., 2024), while OnlineHMR explores online world-grounded motion recovery (Zhao et al., 2026). Multi-person reconstruction raises a further question, since a per-person crop does not scale with the number of people. Bottom-up regressors answer it in a single pass (Baradel et al., 2024). Joint human-scene reconstruction instead treats the people and the scene as one problem. Scene constraints help resolve human pose ambiguities (Hassan et al., 2019), and joint optimization estimates human geometry, scene structure, and camera poses from multi-view images (Müller et al., 2025) or monocular videos with human-scene contact constraints (Liu et al., 2026). HAMSt3R incorporates human information into feed-forward scene reconstruction from multi-view images (Rojas et al., 2025). Human3R jointly predicts human meshes, scene geometry, and camera poses in a recurrent model, enabling online reconstruction of multiple people and their surroundings (Chen et al., 2026b).

Streaming 3D reconstruction. Pointmap regression supports uncalibrated reconstruction (Wang et al., 2024a) and geometry-aware matching (Leroy et al., 2024), and extends to dynamic scenes (Zhang et al., 2025a). VGGT jointly predicts camera parameters, pointmaps, depth, and point tracks in a unified model (Wang et al., 2025a), but processes a fixed set of frames at once. Incremental methods instead carry information forward, and differ in what they keep and how they update it. Spann3R keeps a spatial memory and predicts globally aligned pointmaps without optimization-based global alignment (Wang & Agapito, 2024), keeping the map consistent as frames are added. SLAM3R combines local reconstruction with learned global registration, which limits drift over long sequences (Liu et al., 2025). CUT3R integrates historical observations through a recurrent state (Wang et al., 2025b), whose contents summarize everything seen so far and condition each new frame. StreamVGGT and STream3R instead use causal attention over cached tokens (Zhuo et al., 2025; Lan et al., 2026). WinT3R combines local windows with a camera token pool to retain broader spatial context (Li et al., 2026). TTT3R adapts state updates to balance historical information and new observations (Chen et al., 2026a). TTT3R and Human3R additionally reset the state at fixed intervals and align the resulting chunks by their cameras, which extends inference beyond the sequence length seen in training and bounds the drift that the state would otherwise accumulate (Chen et al., 2026a;b).

Multi-shot human reconstruction. Multi-shot reconstruction addresses the discontinuities that editing introduces. A shot change replaces the viewpoint without breaking the recorded timeline, so the frames on either side of it are adjacent in time yet unrelated in appearance. This is what distinguishes the setting from reconstruction on a single continuous take. Pavlakos et al. (2022) exploit observations across shot changes as complementary views for human mesh recovery. HumanMM (Zhang et al., 2025b) extends this direction to world-frame motion recovery for a single person, aligning shots and integrating motion over the sequence. ShowMak3r (Kim et al., 2025) reconstructs actors and their surroundings from television footage while tracking identities across shots. For multiple people, Multi-THuMBS (On et al., 2026) registers the boundary frames in a shared 3D space before matching, combines geometric, appearance, and pose cues for identity association, and refines trajectories afterwards through temporal smoothing and cross-camera constraints. These works treat a shot change either as a multi-view problem, in which the boundary frames supply complementary views of one moment, or as a tracking problem, in which identity has to be carried across it. Where identity has to be established across a shot change, these methods draw on appearance. A shot change renders that cue unreliable, since the viewpoint changes

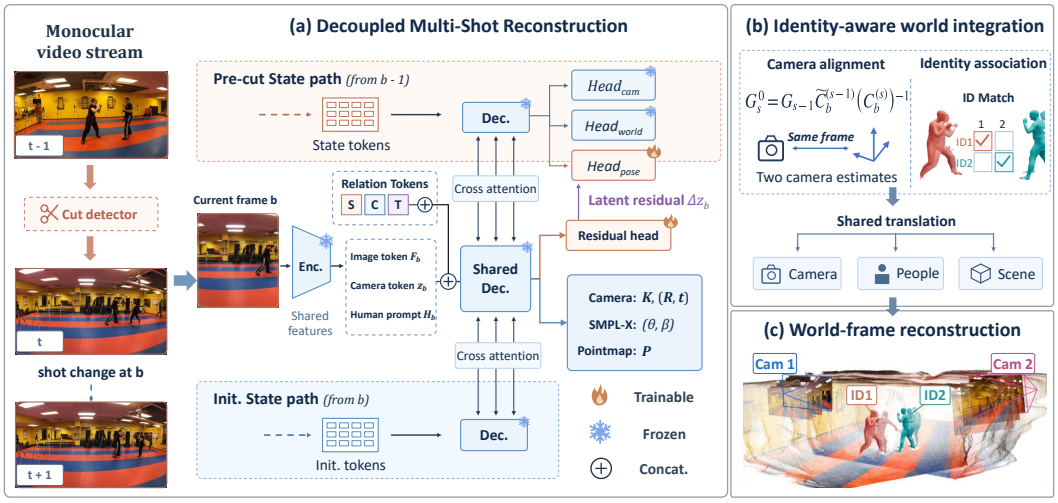

Figure 3: **Method overview.** (a) A cut detector marks each shot change, and the boundary frame is used twice, along a history-conditioned path and after a state reset; relation tokens read its features together with the accumulated history. (b) Camera alignment and identity association place the new shot in the shared world frame. (c) The two shots are then reconstructed together in that frame, with their cameras, the people, and the scene points.

abruptly while the body configuration does not. Our association therefore uses predicted geometry alone and no appearance features.

3 Method

We consider streaming 4D reconstruction of people and scene from multi-shot monocular video. At time t, the goal is to recover camera poses, multi-person body geometry, and scene pointmaps in a shared world frame, so that the same person keeps one location across shots. Inference is forward-only, so the estimate at time t may use only $I_1, \ldots, I_t$. Below, $S_t^{(s)} \in \mathbb{R}^{768\times768}$ denotes the recurrent state, a set of tokens the base model maintains. The camera pose of frame t in that shot's coordinates is the homogeneous matrix $C_t^{(s)} \in \mathbb{R}^{4\times4}$, and $G_s \in \mathbb{R}^{4\times4}$ maps the shot into the shared world frame. Throughout, the superscript (s) marks a quantity of shot s, with shots numbered in temporal order, the subscript t a time step, and the shot transform carries the same index as a subscript, G_s. A shot change happens at one such step, which we write b. The boundary frame is frame b, and frame $b-1$ is the last frame of the preceding shot. *History* denotes the state accumulated from earlier frames, together with the features derived from it. Sections 3.1 to 3.3 decouple cross-shot reference from temporal recurrence, estimate the inter-shot relation, and establish identities in the shared world frame. Figure 3 provides an overview.

3.1 Decoupled Multi-Shot Reconstruction

Shot3R carries history in a recurrent state that determines how each frame is reconstructed, inherited from the base model Human3R (Chen et al., 2026b). It is designed for a continuous take and does not consider multi-shot video. At a shot change the features the state has accumulated were formed from the preceding shot rather than the new one, so carrying that state forward degrades the new shot's reconstruction, most when the viewpoint change is large; reconstructing the new shot without it is no better, since two independently reconstructed shots define separate world frames with no correspondence between them (Section 4.3 tests this case directly). To tell when a shot change occurs, Shot3R adds a cut detector over RGB frame pairs. The frame it flags is the boundary frame b. Frame b is special, being both the frame at which the viewpoint jumps and the first frame of the new shot. It therefore belongs to the new shot while remaining comparable with the pre-cut state, so Shot3R uses it in two roles, the pre-cut and init state paths of Figure 3(a).

The first is the learned history-conditioned path $\mathcal{A}_\theta$, which reads the pre-cut state $S_{b-1}^{(s-1)}$ and returns the pose of the new view expressed in the coordinates of the preceding shot; the second discards that history and reconstructs the frame under the new viewpoint from a state initialized by the frame itself, $S_{\text{init}}(I_b)$,

$$\begin{aligned} \widetilde{C}_b^{(s-1)} &= \mathcal{A}_\theta(I_b, S_{b-1}^{(s-1)}), \\ S_b^{(s)} &= \text{Human3R}(I_b, S_{\text{init}}(I_b)). \end{aligned} \tag{1}$$

That reconstruction carries the boundary frame's camera pose in the new shot's own coordinates, $C_b^{(s)}$. Since both paths estimate the pose of this same frame, the frame becomes the anchor between the two shots, and the two estimates must agree once both are expressed in the world frame,

$$G_{s-1}\, \widetilde{C}_b^{(s-1)} = G_s^0\, C_b^{(s)}. \tag{2}$$

Solving Eq. (2) for the new shot gives the mapping G_s^0, the camera-derived initial estimate of G_s, and G_{s-1} is the transform already established for the preceding shot. The mapping between the two shots therefore follows in closed form from two pose estimates alone, without correspondence search or joint optimization. Only $S_b^{(s)}$ continues beyond the boundary; the state maintained along the history-conditioned path serves solely to relate the two views.

3.2 History-Conditioned Spatial Reasoning

The relation between the two shots is estimated from three relation tokens and converted into an explicit transformation. The base model carries the evidence for this relation in its state and memory, but it stays implicit and reaches the new shot only through the recurrent update. We make it explicit as a correction module that combines the three tokens into one relation, which the boundary step then applies directly. Below, $\bar{X}$ is the mean of a token set X over its token axis, taken as zero when empty, and every token and latent here has the base model's width of $\mathbb{R}^{768}$.

Semantic relation token. The image tokens F_b, the camera token z_b, and the human prompts $\{H_{b,j}\}$ describe the current observation, while the base model's persistent scene state and camera memory carry evidence from history. These two sides are the main evidence for the relation at the boundary. Each side is summarized by cross-attention with a learned relation query, giving r_b^{cur} and r_b^{mem}, and a learned gate combines the two,

$$\begin{aligned} r_b^{\text{sem}} &= \gamma_b \odot r_b^{\text{cur}} + (1 - \gamma_b) \odot r_b^{\text{mem}}, \\ \gamma_b &= \sigma\big(\text{MLP}([\, r_b^{\text{cur}},\ r_b^{\text{mem}}\,])\big), \end{aligned} \tag{3}$$

where $\odot$ is the elementwise product, $[\cdot]$ concatenates along the token axis, MLP is a two-layer perceptron, r_b^{mem} additionally reads the relation tokens cached at the previous boundary, and σ is the sigmoid function. The last layer of the gate is zero-initialized, so it starts as an equal mix of the two and stays usable when history is unavailable.

Camera relation token. Because the relation we estimate is a camera relation, the change in the camera token across the shot change is its most direct evidence. We therefore form

$$r_b^{\text{cam}} = \text{MLP}\big([\, z_b,\ \bar{z}_{b-1},\ z_b - \bar{z}_{b-1},\ r_b^{\text{mem}}\,]\big), \tag{4}$$

where $\bar{z}_{b-1}$ is the camera token of the preceding frame averaged over its tokens and r_b^{mem} is the memory side of the semantic relation token.

Temporal relation token. The previous boundary shows how a comparable shot change was resolved, which is a prior for the current one, but it is not always present; the token therefore encodes the relation tokens, latent residuals, and gate values cached at the previous boundary,

$$r_b^{\text{tmp}} = \text{MLP}\big([\, \overline{\mathcal{R}}_{b'},\ \overline{\Delta z}_{b'},\ g_{b'}\mathbf{1}\,]\big), \tag{5}$$

with $\Delta z_{b'}$ and $g_{b'}$ the latent residual and gate value applied at the previous boundary b', the bar denoting the mean over tokens and $\mathbf{1}$ the all-ones vector that broadcasts the gate over the token axis. The same cached tokens also feed r_b^{mem}. Each token is summed with its own learned type embedding e_i, and the three are inserted immediately after the camera token, so that the shared

decoder refines them together with the image tokens and the human prompts, conditioned on the pre-cut state $S_{b-1}^{(s-1)}$,

$$\mathcal{R}_b = \big[\, r_b^{\mathrm{sem}} + e_1,\; r_b^{\mathrm{cam}} + e_2,\; r_b^{\mathrm{tmp}} + e_3 \,\big], \qquad \widehat{\mathcal{R}}_b = \mathrm{Dec}\big(\big[\, z_b,\; \mathcal{R}_b,\; F_b,\; \{H_{b,j}\} \,\big],\; S_{b-1}^{(s-1)}\big). \tag{6}$$

The tokens $\widehat{\mathcal{R}}_b$ decoded in Eq. (6) are mean-pooled over the token axis into a shared relation feature ρ_b that a residual head turns into a latent residual, gated by a shared gate g_b. The relation therefore reaches the reconstruction as one explicit correction, applied to the latent from which the pose head reads the camera estimate. For the decoded camera latent c_b, the update is

$$\begin{aligned} g_b &= \sigma(f_g(\rho_b)), \\ \widehat{c}_b &= c_b + g_b\, \Delta z_b, \end{aligned} \tag{7}$$

where f_g and f_c are the residual head's learned mappings and $\Delta z_b = f_c(\rho_b)$ is the latent residual that the gate scales. The correction reaches only the pose head, which supplies the history-conditioned camera estimate $\widetilde{C}_b^{(s-1)}$ required by Eq. (2); the other two prediction heads of Figure 3(a), $\mathrm{Head}_{\mathrm{cam}}$ and $\mathrm{Head}_{\mathrm{world}}$, read the decoded latents. G_s^0 then follows from Eq. (2) and places the new shot's local reconstruction in the shared world frame, providing the initial geometry for identity association and spatial refinement.

3.3 Identity-Aware World Integration

The camera-derived mapping G_s^0 brings both sides of the shot change into the shared world frame, so a person's position can be compared across the change. The two frames are also adjacent in time, because a shot change never interrupts the recorded timeline, and a person's configuration therefore changes little between them. Identity can then be established from body configuration, which the base model provides as joint positions and body meshes, without any appearance model. Human mesh recovery represents pose relative to the root (Kanazawa et al., 2018; Kolotouros et al., 2019), and the pelvis and torso tend to change least between two adjacent frames, while the limbs and hands may move substantially. We compare pelvis position, torso orientation, and the root-centered joint configuration through a normalized pairwise cost, and Hungarian matching yields the one-to-one match set $\mathcal{M}_b$. The matcher does not reject matches, so an assignment can be geometrically plausible yet incorrect (Appendix B.3).

Human matches also refine the shared spatial relation. For a matched pair, the displacement between the pre-cut pelvis p_i^- and the new-shot pelvis p_j^+ after the initial mapping measures the residual misalignment that the camera-derived mapping has not absorbed, and its componentwise median is robust to individual mismatches. We take this median as a shared translation, scaled by a constant α,

$$\Delta_b = \alpha \operatorname*{median}_{(i,j)\in\mathcal{M}_b} \big(p_i^- - p_j^+\big), \qquad G_s = \mathrm{Trans}(\Delta_b)\, G_s^0. \tag{8}$$

where $\mathrm{Trans}(\Delta_b)$ in Eq. (8) is the corresponding homogeneous translation. When no valid match is available, the camera-derived mapping is retained unchanged, and applying one transformation to cameras, humans, and scene points preserves their relative geometry within the shot.

The shot transform G_s remains fixed within the shot. Each subsequent camera pose is expressed in the world frame as $C_t^w = G_s C_t^{(s)}$, and each homogeneous 3D point of the reconstruction by the same mapping. Matched people inherit existing identities, while unmatched new person slots receive new identities, and within the shot identities follow the base model's person slots. At the next shot change, the current world representation provides the reference for the next boundary step.

The three components above are trained together. The pretrained reconstructor remains frozen, and only the newly added modules and the head adapters (Hu et al., 2022) are updated (Appendix B.1).

4 Experiments

The experiments test the insight the method is built on, that history can inform the cross-shot relation without conditioning the new shot's reconstruction.

Table 1: **Multi-shot reconstruction.** Results on EgoBody and EgoHumans, with EgoBody support in parentheses (Table 5 for EgoHumans). Shot3R is best on every geometry and camera column of both datasets, and on identity on EgoBody; bold marks the best value and shading the three best.

		EgoBody ($N = 129$)					**EgoHumans** ($N = 90$)				
Method	Online	W ↓	WA ↓	ATE ↓	IDF1 ↑	Coverage ↑	W ↓	WA ↓	ATE ↓	IDF1 ↑	Coverage ↑
JOSH (127)	×	399.6	322.9	0.087	0.973	0.980	1410.7	481.3	1.755	**0.616**	**0.686**
OnlineHMR (97)	×	961.2	418.1	0.259	0.375	0.424	1311.7	405.5	3.869	0.142	0.211
Human3R (129)	✓	321.1	233.4	0.124	0.870	**0.981**	1005.1	365.7	1.673	0.471	0.639
Shot3R (129)	✓	**258.8**	**201.7**	**0.016**	**0.985**	**0.981**	**866.8**	**336.9**	**1.423**	0.571	0.616

Datasets and evaluation. We build sequential monocular streams from EgoBody (Zhang et al., 2022), EgoHumans (Khirodkar et al., 2023), and Harmony4D (Khirodkar et al., 2024) by introducing shot changes into their multi-view recordings, supplying one view at a time, and we also evaluate an edited film clip in Section 4.2. Human reconstruction is measured in millimeters by W-MPJPE (W), which aligns the initial frame, and WA-MPJPE (WA), which aligns all matched observations. Camera accuracy is measured in meters by ATE, with Sim(3) alignment on EgoBody and Harmony4D and SE(3) on EgoHumans; identity consistency is measured by IDF1 and Coverage, the fraction of visible ground-truth person-frames matched to predictions. Methods are compared only within a dataset, and geometry errors are reported on the valid support each method provides. Appendix A gives the protocol construction, metric conventions, and exclusions.

Baselines and implementation. We compare Shot3R with Human3R (Chen et al., 2026b), OnlineHMR (Zhao et al., 2026), JOSH (Liu et al., 2026), TRACE (Sun et al., 2023), and PromptHMR (Wang et al., 2025c) with SPEC (Kocabas et al., 2021). Because the methods differ in their outputs and inputs, we compare each on the outputs it supports, reporting clip support for the partial ones (Appendix C.4); a negative Online entry marks a pipeline that uses the complete sequence or global post-processing. JOSH runs offline, our OnlineHMR runs use its global-camera backend, and continuous-state Human3R is the controlled reference built on the same reconstructor. HumanMM (Zhang et al., 2025b) and Multi-THuMBS (On et al., 2026) were not available as runnable implementations at the time of evaluation and are discussed as related work only. Only the shot-change modules and head adapters are trained, on AvatarReX, THuman, MVHuman100/200, and MultiHuman (Appendix B).

4.1 Multi-Shot Human-Scene Reconstruction

Table 1 reports multi-shot reconstruction on EgoBody and EgoHumans. The clearest margins are on the two quantities the boundary step produces directly, camera placement and person identity. Camera ATE falls by **87%** on EgoBody against the continuous-state reference, and IDF1 reaches **0.985**. Among the baselines, JOSH trades placement for identity. Its offline optimization matches

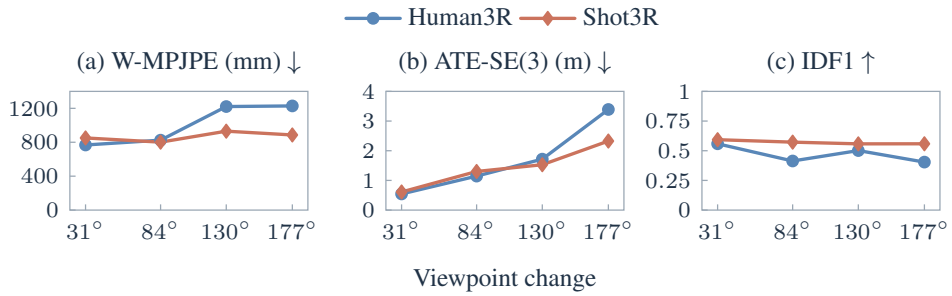

Figure 4: **Human, camera, and identity performance across viewpoint changes.** W-MPJPE, ATE-SE(3), and IDF1 on the four predefined EgoHumans strata, containing 24, 22, 22, and 22 clips. Points are clip-averaged. WA and Coverage are in Appendix C.2.

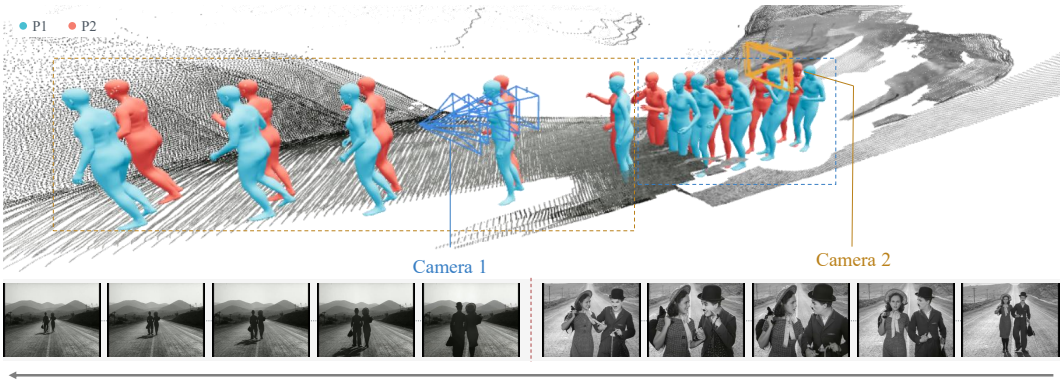

Figure 5: **Qualitative reconstruction on film footage.** Reconstruction across a frontal-to-rear shot change, with cyan and coral for the two people and blue and gold for the pre- and post-cut cameras. The input strip runs right to left, the dashed line marking the shot change.

people most completely, giving the highest IDF1 and Coverage on EgoHumans, but the largest human errors. Harmony4D results are in Appendix C.1.

Shared-world alignment. We test whether one shared transform keeps cameras and people coherent after the shot change, on the EgoHumans captures (Appendix C.3). Reinitializing without relating the two views is worse than the continuous-state baseline on every quantity, and a shared alignment alone leaves the post-cut root error above it. Under that alignment the full system lowers root error to **2.69 m** and camera translation to **4.55 m**, both below the baseline. The alignment carries the cross-shot relation, the continued state does not.

Efficiency. On five fixed 150-frame streams on an NVIDIA L20, Shot3R runs **5.7×** faster than JOSH at **1.4×** the cost of the continuous-state reference (Appendix E.1).

4.2 Performance across Shot Changes

Viewpoint changes. Figure 4 reports W-MPJPE, ATE, and IDF1 on the four predefined EgoHumans strata. IDF1 is higher than the continuous-state reference in every stratum, and the geometric advantage grows with the magnitude of the viewpoint change. The propagated recurrent state is formed under the view that has just ended, so the more abrupt the change, the less that state describes the new view and the more the boundary estimate has to contribute. In the largest-change stratum the geometric errors fall by about **30%** relative to that reference. The gains are not uniform, and in the smallest-change stratum both geometric errors are higher for Shot3R; Coverage is slightly lower in all four strata. Appendix C.2 gives the full strata, WA, Coverage, and EgoBody's farthest-view results.

Qualitative comparison. Figure 5 shows Shot3R on an edited film clip. Despite a frontal-to-rear shot change, the reconstruction places the two people and both camera groups in a shared world frame with consistent identities. Figure 6 compares against JOSH and Human3R on three large viewpoint changes, where the baselines misplace the second camera, scatter the people, or assign identities wrongly. The visible failures are the two quantities a shot change puts at risk.

4.3 Controlled Comparisons and Diagnostics

Recurrent-state propagation. We first isolate the recurrent state, holding the checkpoint and the post-shot-change inputs fixed and disabling the Shot3R modules. Over the 16 frames after the shot change, reinitializing the state reduces camera translation and rotation drift by **98%**, with capture-level paired intervals that exclude zero and a one-sided Wilcoxon $p = \mathbf{7.45} \times \mathbf{10^{-9}}$. Nothing else differs between the two settings, so the drift follows from the propagated state rather than from the reconstruction. The measured quantity and the drift curves are in Appendix D.2.

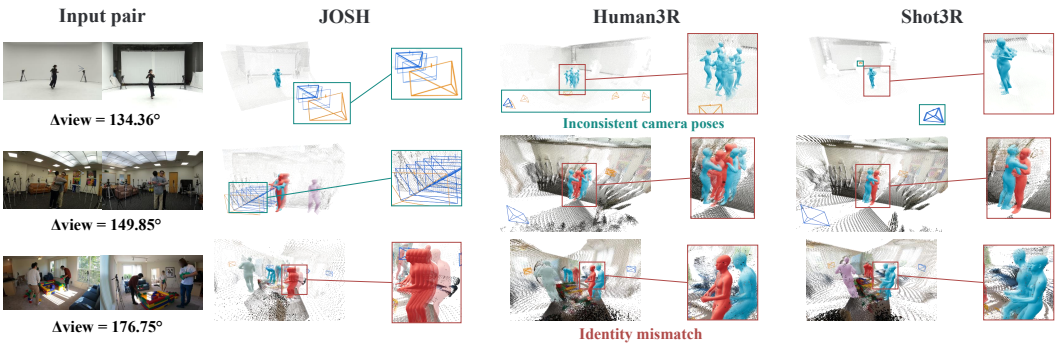

Figure 6: **Qualitative comparison with JOSH and Human3R.** Three comparisons on large viewpoint changes, with insets showing details. Red and teal boxes mark human and camera regions.

Table 2: **Alignment and state reset.** Controls on EgoBody; bold marks the column best. Reset × keeps the previous state; rows 1–4 omit identity association and shared translation.

Alignment	Reset	W ↓	ATE ↓	IDF1 ↑
×	×	321.1	0.124	0.870
×	✓	752.6	1.141	0.834
✓	×	293.3	0.082	0.868
✓	✓	352.8	0.017	0.834
Identity association		349.2	0.019	0.983
Shot3R (full)		**258.8**	**0.016**	**0.985**

Table 3: **Registration alternatives** against Shot3R; bold marks the column best. Human-joint fitting uses shot-change predictions, scene fitting complete shots.

Method	W ↓	ATE ↓	IDF1 ↑
	EgoBody		
Human-joint SE(3)	261.5	0.019	0.896
Scene FPFH+ICP	733.2	0.020	0.834
Shot3R	**258.8**	**0.016**	**0.985**
	EgoHumans		
Human-joint SE(3)	897.0	1.731	0.510
Scene FPFH+ICP	1372.8	3.847	0.455
Shot3R	**866.8**	**1.423**	**0.571**

Alignment and state reinitialization. Table 2 separates the two mechanisms, with identity association and shared translation disabled in rows 1–4. Reinitializing without alignment is worse than leaving the state untouched, because the new shot then loses its link to the shared world frame. With alignment fixed, the reset improves camera ATE while W-MPJPE becomes worse. A better camera estimate alone therefore does not place people correctly, and the two requirements are not interchangeable. Identity association is what moves IDF1, from **0.834** to **0.983**, and it leaves human error largely unchanged. The full system combines all three, and the same separation holds on the full EgoHumans evaluation (Appendix D.3).

Geometric registration alternatives. Table 3 compares Shot3R with two registration alternatives applied to the same independently reconstructed shots. Human-joint SE(3) fitting matches Shot3R in W-MPJPE on EgoBody, and the paired difference includes zero, so the strong control is indistinguishable on that metric. Shot3R obtains lower camera ATE and higher IDF1 on both datasets, so the two are not equivalent even where the human error agrees. The fitting protocol, failed fits, and the remaining controls are in Appendix D. These controls use annotated shot changes, whereas the full system runs the cut detector at inference. On held-out frame pairs that detector reaches F1 **0.982** and false-positive rate **0.031**, and it processes the boundary at its first threshold crossing.

5 Conclusion

We presented Shot3R, the first streaming framework for 4D human-scene reconstruction from multi-shot video. It resolves each shot change from the observed prefix alone, decoupling the two roles of history and preserving identities and alignment through geometric correspondences. Human reconstruction, camera placement, and identity association all improve over the continuous-state reference on EgoBody and EgoHumans, most under large viewpoint changes.

AI USE STATEMENT

We used generative AI tools for writing assistance and polishing, including grammar checking, wording and sentence rewriting, academic phrasing, and shortening redundant text. The authors reviewed and revised all AI-modified content.

We used generative AI tools for retrieval and discovery of literature and research directions related to this topic, including suggested search terms and overviews of related lines of work. This output served only as a search aid, and the authors verified the existence, the content, and the relevance of every reference cited in this paper.

We used generative AI tools for research ideation and execution, namely discussions of research ideas and of method and experiment design, implementation and debugging, and analysis of results. The authors made the final decisions on the research design, the choice of method, the execution of the experiments, and the interpretation of the results.

We used generative AI tools to draft portions of the paper from research content, points, outlines, methods, and experimental results supplied by the authors. The authors checked, revised, rewrote, and integrated this text, and no AI-generated content was used without review. The authors take full responsibility for the research content, the references, the experimental results, and the final text of this paper.

REPRODUCIBILITY STATEMENT

Section 3 specifies the two inference paths, spatial mappings, and identity integration. Appendix A describes stream construction, alignment conventions, aggregation, and valid support. Appendix B documents the learned modules, the cut detector, and state and identity handling. Appendices C–E report extended results, ablations and diagnostics, and timing protocols.

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. URL https://arxiv.org/abs/2402.14654.

Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791. URL https://doi.org/10.1109/34.121791.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/530eb593743559fa3c926a1fc1f520e9-Abstract-Conference.html.

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*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/467c017e2639255e9c5d91a2e582f95a-Abstract-Conference.html.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692.

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*, pp. 14783–14794, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Goel_Humans_in_4D_Reconstructing_and_Tracking_Humans_with_Transformers_ICCV_2023_paper.html.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2282–2292, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Hassan_Resolving_3D_Human_Pose_Ambiguities_With_3D_Scene_Constraints_ICCV_2019_paper.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1712.06584.

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. EgoHumans: An egocentric 3D multi-human benchmark. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2305.16487.

Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: A video dataset for in-the-wild close human interactions. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2410.20294.

Sangmin Kim, Seunguk Do, and Jaesik Park. ShowMak3r: Compositional TV show reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kim_ShowMak3r_Compositional_TV_Show_Reconstruction_CVPR_2025_paper.html.

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*, pp. 5253–5263, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Kocabas_VIBE_Video_Inference_for_Human_Body_Pose_and_Shape_Estimation_CVPR_2020_paper.html.

Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *IEEE/CVF International Conference on Computer Vision*, pp. 11035–11045, 2021. URL https://arxiv.org/abs/2110.00620.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2252–2261, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Kolotouros_Learning_to_Reconstruct_3D_Human_Pose_and_Shape_via_Model-Fitting_ICCV_2019_paper.html.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109. URL https://doi.org/10.1002/nav.3800020109.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.10893.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73220-1_5. URL https://arxiv.org/abs/2406.09756.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2101.08779.

Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. WinT3R: Window-based streaming reconstruction with camera token pool. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/617ae68c72b8bd8aeb4148eee81bc0c8-Abstract-Conference.html.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16651–16662, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_SLAM3R_Real-Time_Dense_Scene_Reconstruction_from_Monocular_RGB_Videos_CVPR_2025_paper.html.

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. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fc657b7fd7b9aaa462f2ef9f0362b273-Abstract-Conference.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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*, pp. 21948–21958, 2025. URL https://arxiv.org/abs/2412.17806.

Jeongwan On, Muhammad Salman Ali, Muneeb A. Khan, Sunwoo Park, Inwoong Moon, Hyung Jin Chang, Jaekwang Kim, Seong Jong Ha, and Seungryul Baek. Multi-THuMBS: Multi-person tracking of 3D human meshes beyond video shots. *arXiv preprint arXiv:2607.01626*, 2026. URL https://arxiv.org/abs/2607.01626.

Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Pavlakos_Human_Mesh_Recovery_From_Multiple_Shots_CVPR_2022_paper.html.

Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshops*, 2016. URL https://arxiv.org/abs/1609.01775.

Sara Rojas, Matthieu Armando, Bernard Ghanem, Philippe Weinzaepfel, Vincent Leroy, and Gregory Rogez. HAMSt3R: Human-aware multi-view stereo 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2508.16433.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision*, 2011. doi: 10.1109/ICCV.2011.6126544. URL https://ieeexplore.ieee.org/abstract/document/6126544.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation*, pp. 3212–3217, 2009. doi: 10.1109/ROBOT.2009.5152473. URL https://ieeexplore.ieee.org/abstract/document/5152473.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. doi: 10.1145/3680528.3687565. URL https://doi.org/10.1145/3680528.3687565.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.07531.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012. doi: 10.1109/IROS.2012.6385773. URL https://ieeexplore.ieee.org/abstract/document/6385773.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2306.02850.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. URL https://arxiv.org/abs/2408.16061.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. URL https://arxiv.org/abs/2503.11651.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://arxiv.org/abs/2501.12387.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a. URL https://arxiv.org/abs/2312.14132.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *European Conference on Computer Vision*, 2024b. URL https://arxiv.org/abs/2403.17346.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c. URL https://arxiv.org/abs/2504.06397.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2302.12827.

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*, pp. 11038–11049, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yuan_GLAMR_Global_Occlusion-Aware_Human_Mesh_Recovery_With_Dynamic_Cameras_CVPR_2022_paper.html.

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*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce1710ff5b66f89b0d707bf1ec2850bf-Abstract-Conference.html.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2112.07642.

Yuhong Zhang, Guanlin Wu, Ling-Hao Chen, Zhuokai Zhao, Jing Lin, Xiaoke Jiang, Jiamin Wu, Zhuoheng Li, Hao Frank Yang, Haoqian Wang, and Lei Zhang. HumanMM: Global human motion recovery from multi-shot videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_HumanMM_Global_Human_Motion_Recovery_from_Multi-shot_Videos_CVPR_2025_paper.html.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and Laszlo A. Jeni. OnlineHMR: Video-based online world-grounded human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.17355.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. URL https://arxiv.org/abs/2507.11539v2. Version 2, 2026.

APPENDIX

TABLE OF CONTENTS

A Evaluation Protocols . . . 16
A.1 Monocular streams and evaluation sets . . . 16
A.2 Metrics and alignment conventions . . . 16
A.3 Aggregation, valid support, and baselines . . . 16

B Implementation Details . . . 17
B.1 Learned modules and training . . . 17
B.2 Cut detector . . . 17
B.3 State, coordinates, and identity handling . . . 18

C Extended Results . . . 18
C.1 Harmony4D . . . 18
C.2 Effect of viewpoint change . . . 18
C.3 Shared-world alignment . . . 18
C.4 External methods with partial geometry support . . . 19
C.5 Three-shot reconstruction . . . 20

D Ablations and Diagnostics . . . 21
D.1 Alignment, state propagation, and registration . . . 21
D.2 Camera drift under state propagation . . . 24
D.3 System-level ablation . . . 24

E Efficiency . . . 25
E.1 Complete prediction pipelines . . . 25
E.2 Effect of shot-change count . . . 25

F Scope and Limitations . . . 26

G Qualitative 4D human-scene reconstruction results . . . 26
G.1 Outdoor sequence . . . 26
G.2 Reconstruction examples across datasets . . . 26

A Evaluation Protocols

A.1 Monocular streams and evaluation sets

We construct sequential monocular streams from synchronized captures in EgoBody, EgoHumans, and Harmony4D. Each stream observes one camera at a time and switches viewpoints while the underlying timeline advances. No simultaneous second view is supplied. Table 4 summarizes the evaluation sets used throughout the paper. Camera calibration, human annotations, and identities are used by the evaluator and are not inputs to Shot3R. Annotated shot changes are supplied only to controls explicitly identified as such.

Table 4: **Multi-shot evaluation protocols.** Each clip contains two sequential monocular shots.

Dataset	Clips	Captures	Frames / clip	Aggregation
EgoBody	129	43	150	Recording mean
EgoHumans	90	27	100	Clip mean
Harmony4D	88	25	150	Clip mean

The retained EgoHumans set contains 90 of 116 inferred camera-pair streams, and Harmony4D contains 88 of 100 candidate streams. Every retained stream satisfies the eligibility condition of the frozen protocol. On EgoHumans the base model establishes a valid human anchor before the shot change, and on Harmony4D an initial prediction-to-ground-truth match exists for the shared world fit; the remaining 26 and 12 streams satisfy neither condition. The same retained set and the same RGB inputs are used for every compared method, and additional method-specific unavailability is reported separately.

A.2 Metrics and alignment conventions

W-MPJPE (W) measures world-frame joint error under a fixed initial Sim(3) alignment, capturing deviations from the initial spatial reference. WA-MPJPE (WA) measures joint error after alignment over all matched observations. On EgoBody, one transform is shared by all matched people in a clip. W uses the earliest two valid pre-shot-change time points, and WA uses all valid time points. On EgoHumans and Harmony4D, the evaluator fits W and WA separately for each matched identity track, using its first two available observations and its complete matched trajectory, respectively. These per-track fits can absorb relative placement errors between people. They should therefore be interpreted together with camera and identity metrics, and absolute W/WA values should not be compared across datasets.

Camera ATE (Sturm et al., 2012) measures global camera trajectory accuracy through the root-mean-square error of camera centers after trajectory alignment. Sim(3) alignment applies to EgoBody and Harmony4D; EgoHumans uses SE(3). Human errors are in millimeters and camera errors in meters. IDF1 (Ristani et al., 2016) measures identity consistency through sequence-level identity association. Coverage counts the visible ground-truth person-frames for which the evaluator accepts a prediction match, as a fraction of all such frames. It measures represented observations, rather than geometric accuracy. Consequently, a lower geometry error with lower Coverage does not establish an improvement on all people in the stream.

A.3 Aggregation, valid support, and baselines

EgoBody first averages valid clips within each recording and then weights its 43 recordings equally. The headline EgoHumans and Harmony4D results average valid clips. Geometry errors are conditional on the support reported in Table 5, and methods are compared on the outputs each one supports. IDF1 and Coverage retain the complete denominator of each reported evaluation set. The state diagnostic aggregates over captures equally, and EgoHumans intervals resample captures before clips, to account for correlated camera pairs.

We use the released implementations of the external methods. JOSH processes complete sequences offline with Pi3X initialization; we supply no ground-truth camera, identity, body shape, shot change, or shot-wise transform. The online flags in Table 1 describe the evaluated pipeline rather

Table 5: **Valid geometry support.** Entries give the number of clips with W / WA / ATE, respectively. A dash denotes an unavailable camera output under this protocol.

Method	EgoBody (129)	EgoHumans (90)	Harmony4D (88)
PromptHMR (SPEC)	64 / 87 / 129	17 / 25 / –	5 / 26 / 76
JOSH	127 / 127 / 127	77 / 77 / 87	88 / 88 / 88
OnlineHMR	97 / 97 / 129	50 / 50 / 87	30 / 30 / 88
TRACE	22 / 35 / –	43 / 59 / –	4 / 26 / –
Human3R	129 / 129 / 129	90 / 90 / 90	88 / 88 / 88
Shot3R	129 / 129 / 129	90 / 90 / 90	88 / 88 / 88

than all modes of the original method. TRACE and PromptHMR (SPEC) mainly contribute human and identity results, and their geometric outputs cover only a subset of clips; PromptHMR additionally supplies an evaluable camera trajectory on EgoBody and Harmony4D, the two sets that use Sim(3) alignment, reported with its support in Appendix C.4 and Table 5. HumanMM and Multi-THuMBS are treated as related work only; their published numbers are not mixed with our protocol. TRAM is included in the runtime study and the qualitative comparisons in Appendix C.2.

B Implementation Details

B.1 Learned modules and training

The pretrained base model stays frozen; the trainable parameters are the history-conditioned path, the relation tokens with the residual head, and the LoRA adapters, trained on MultiHuman, MVHuman100/200, THuman, and AvatarReX. One checkpoint and configuration is shared by all three evaluation protocols.

Each adapted linear layer computes $Wx + \frac{\alpha_{\mathrm{LoRA}}}{r} BAx$, where W is the frozen pretrained weight and A, B form a rank-r factorization with $r = 8$ and $\alpha_{\mathrm{LoRA}} = 8$. The up-projection B is zero-initialized, so an adapted layer starts from the pretrained mapping.

Boundary adaptation uses three-view events that contain two pre-cut observations and one post-cut observation, which concentrates supervision on the boundary frame. The primary supervision is the camera rotation and translation of that frame, together with human translation where annotations are available. Auxiliary terms regularize the magnitude of the latent residual and keep the local pointmaps, body parameters, and world pointmaps consistent with respect to predictions produced with boundary adaptation disabled.

B.2 Cut detector

The detector computes 20 statistics from adjacent RGB frames, including color and grayscale changes, perceptual-hash distances, edge changes, optical-flow magnitude statistics, histogram distances, and ORB (Rublee et al., 2011) correspondence statistics. A history of three frame-pair descriptors is passed to a single-layer GRU (Cho et al., 2014) with hidden dimension 32 and a linear binary head. We train it for 120 epochs with binary cross-entropy and AdamW (Loshchilov & Hutter, 2019), using learning rate 2×10^{-3} and weight decay 2×10^{-3}. The inference threshold is fixed at 0.5. Neither future frames nor reconstruction outputs enter this detector.

The leave-one-source-out diagnostic covers four source groups, each contributing 102 held-out frame pairs with 54 positives, for 408 pairs in total. The GRU achieves source-macro F1 of 0.982, false-positive rate of 0.031, and Brier score of 0.015. These statistics describe the constructed detector diagnostic. In the primary two-shot wrapper, the first threshold crossing triggers boundary processing; the experiment does not measure unrestricted shot-change detection over long edited videos.

B.3 State, coordinates, and identity handling

At a shot change, the history-conditioned path reads a copy of the preceding state. Missing relation history is represented by zero inputs. Its inferred camera relation anchors the new shot, but the temporary state is discarded. Temporary human predictions provide training supervision and are not substituted for the bodies emitted by the independent path. Camera poses follow the camera-to-world convention. The shot transform is left-composed with camera poses and applied to human joints, mesh vertices, and scene points, and remains fixed until the next shot change. Previously emitted frames are not revised.

Identity association uses pelvis position, torso orientation, and root-centered joint configuration, with each pairwise cost normalized by the median of its valid values before the terms are added. Hungarian assignment (Kuhn, 1955) returns one-to-one matches among available predictions. The current implementation has no appearance-based re-identification model or additional rejection threshold. Matched slots inherit prior identities, and unmatched new slots receive new identities. If no person pair is available, the shared translation refinement is omitted and camera-based alignment is retained. The translation coefficient remains $\alpha = 0.5$ across datasets.

C Extended Results

C.1 Harmony4D

Table 6 extends the main comparison to Harmony4D. The pattern matches EgoBody: the largest gains are on camera placement and identity, with ATE falling from 0.142 to 0.017 m and IDF1 rising from 0.553 to 0.636 compared with the continuous-state reference. W stays slightly higher, which the per-track fit described in Appendix A.2 can absorb, and JOSH again trades placement for coverage, matching more people than Shot3R but placing them far worse.

Table 6: **Multi-shot reconstruction on Harmony4D.** W and WA are in mm; ATE uses Sim(3) and is in m. Best values are bold. Valid support is given in Table 5.

Method	W ↓	WA ↓	ATE ↓	IDF1 ↑	Coverage ↑
PromptHMR (SPEC)	963.2	306.9	0.478	0.030	0.034
JOSH	770.2	369.7	0.153	0.551	**0.959**
OnlineHMR	750.8	335.9	1.364	0.068	0.159
TRACE	677.1	270.2	–	0.016	0.016
Human3R	**496.6**	259.6	0.142	0.553	0.915
Shot3R	519.6	**247.0**	**0.017**	**0.636**	0.913

C.2 Effect of viewpoint change

Table 7 complements Figure 4 with W, WA, ATE, IDF1, and Coverage on EgoHumans. The four viewpoint groups contain 24, 22, 22, and 22 clips, and entries are clip-macro means within each stratum. At mean changes of 130.4° and 177.0°, Shot3R reduces W, WA, and ATE and increases IDF1 relative to Human3R. At 31.1°, its geometry errors are higher, and Coverage is slightly lower across all four groups.

On EgoBody's farthest-view subset, Shot3R reduces W-MPJPE from 505.1 to 354.3 mm and increases IDF1 from 0.758 to 0.983 relative to Human3R, with unchanged Coverage.

Figures 7–9 expand the comparison in Figure 6 to all seven methods, one dataset per page, at viewpoint changes of 134.36°, 149.85°, and 176.75°.

C.3 Shared-world alignment

This diagnostic tests whether one transform keeps both cameras and people coherent across a shot change. For each clip we estimate one initial Sim(3) fit from the prescribed shared reference and apply that same transform to the complete two-shot prediction, so neither shot receives its own fitted

Table 7: **Multi-shot reconstruction at different viewpoint changes.** W/WA are in mm, ATE-SE(3) is in m, and angles are stratum means. Best values within each viewpoint group are bold.

Angle	Method	W ↓	WA ↓	ATE ↓	IDF1 ↑	Coverage ↑
31.1°	Human3R	**767.7**	**331.5**	**0.544**	0.559	**0.658**
31.1°	**Shot3R**	850.9	357.1	0.610	**0.593**	0.657
84.0°	Human3R	824.7	345.9	**1.145**	0.414	**0.637**
84.0°	**Shot3R**	**801.2**	**341.5**	1.301	**0.572**	0.605
130.4°	Human3R	1221.1	388.6	1.717	0.501	**0.647**
130.4°	**Shot3R**	**930.7**	**331.1**	**1.529**	**0.558**	0.622
177.0°	Human3R	1228.4	400.0	3.390	0.404	**0.611**
177.0°	**Shot3R**	**886.1**	**315.9**	**2.326**	**0.559**	0.576

coordinate system. Table 8 reports post-cut root error, person-pair relative-vector error, and the first-post-cut camera translation and rotation error. Values are averaged within each clip and then equally over captures. Some sequences admit no initial common fit, so per-metric support differs and is listed in Table 9. The continuous-state baseline is available on a different set of captures from the three controlled settings, so the contrast below is indicative rather than support-matched.

Table 8: **Shared-world diagnostic on EgoHumans.** One transform places the complete two-shot prediction in the shared world frame. Distances are in m and camera rotation in degrees; the best value in each column is bold. Support counts are given in Table 9.

Setting	Post-cut root ↓	Pair vector ↓	Camera trans. ↓	Camera rot. ↓
Human3R (continued state)	3.08	1.74	5.00	34.40
Reinitialized state	4.30	4.29	9.65	100.61
Alignment only	3.38	**1.49**	5.17	**22.03**
Shot3R	**2.69**	**1.49**	**4.55**	**22.03**

Table 9: **Support for the shared-world diagnostic.** Entries are capture counts with a valid value for each metric.

Method	Post-cut root	Pair vector	Camera trans.	Camera rot.
Human3R	26	20	27	27
Reinitialized state	25	19	27	27
Alignment only	25	19	27	27
Shot3R	25	19	27	27

Three settings form a controlled sequence. Reinitializing the state alone loses the cross-shot world link, and all four errors increase relative to the continuous-state baseline. Adding the shared alignment alone already brings person-pair relative-vector error and camera rotation error to their final values, while post-cut root error remains above the baseline. Shot3R reduces root error and camera translation error further under the same alignment. Reinitialization is therefore not sufficient on its own, and the shot transform accounts for what the reset does not.

C.4 External methods with partial geometry support

Table 10 reports the geometric results of external methods that are not in the main comparison. TRACE and PromptHMR (SPEC) provide evaluable geometry on a subset of clips, and their valid support is listed alongside. These values are therefore conditional on that support and are not directly comparable with methods evaluated on all clips. Two EgoBody clips are unavailable for JOSH because of failures in its own pipeline.

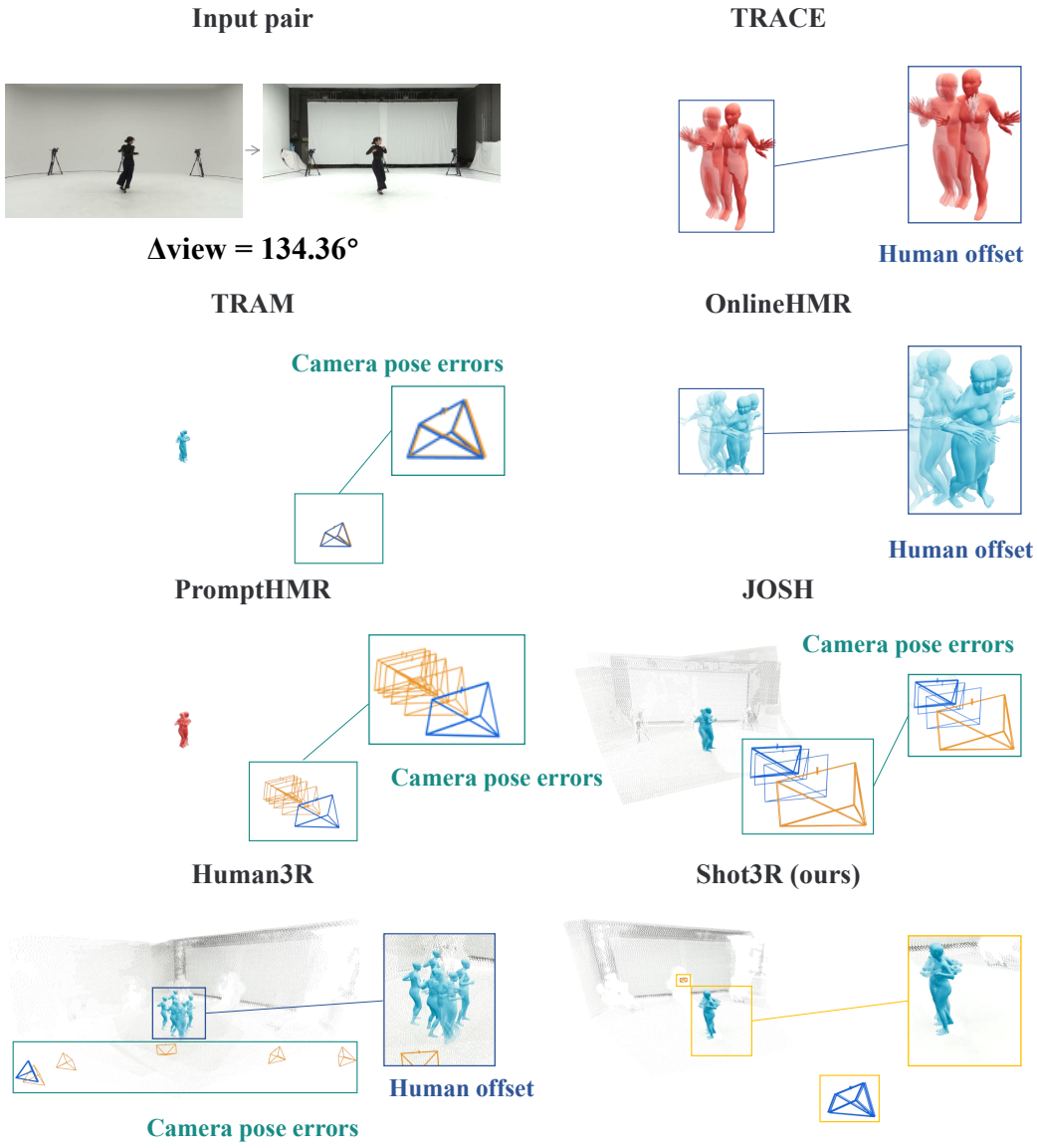

Figure 7: **Full qualitative comparison on AIST++ (Li et al., 2021).** The input pair shows frames before and after a 134.36° viewpoint change. The panels compare TRACE, TRAM, OnlineHMR, PromptHMR, JOSH, Human3R, and Shot3R. Red and teal boxes mark human and camera regions, respectively; connected insets show enlarged details.

C.5 THREE-SHOT RECONSTRUCTION

We further evaluate four Harmony4D clips containing three shots each. Shot3R processes both shot changes sequentially and integrates each new shot into the existing world frame. In Table 11, Shot3R reduces W from 2091.9 to 1304.8 mm and WA from 719.4 to 651.2 mm relative to Human3R, while IDF1 increases from 0.377 to 0.434. ATE increases from 1.618 to 1.700 m. OnlineHMR has valid human errors on only three clips and substantially lower Coverage.

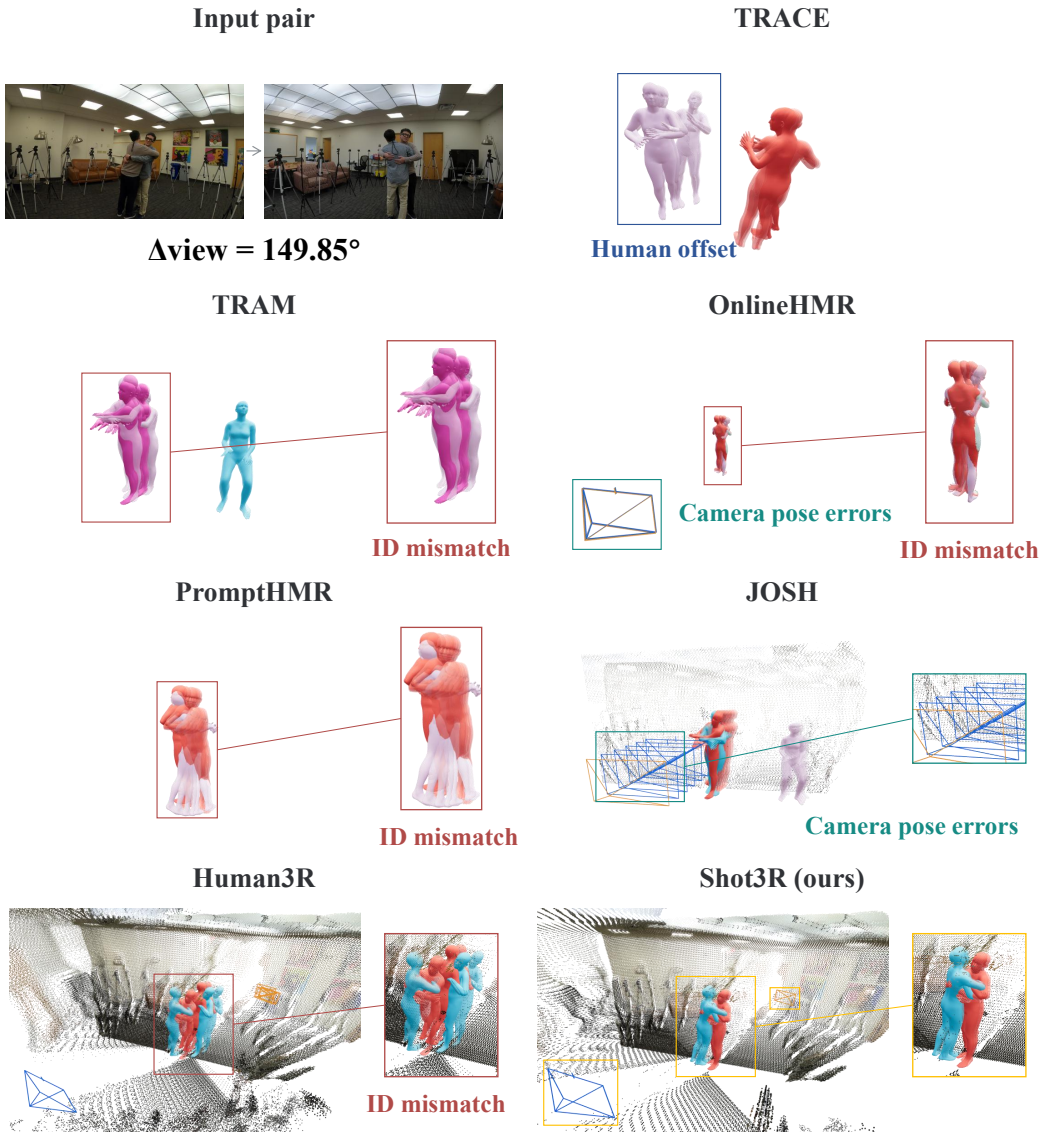

Figure 8: **Full qualitative comparison on EgoHumans.** The input pair spans a 149.85° viewpoint change. The seven methods are shown in the same order as in Figure 7. Full reconstruction views and connected insets display the people and camera configurations, using red for human regions and teal for camera regions.

D Ablations and Diagnostics

D.1 Alignment, state propagation, and registration

Table 12 expands the compact controls in the main text. The two aligned intermediate configurations both omit identity association and shared translation refinement, and differ only in the state propagated after alignment; the association row then adds identity association alone. The boundary operations in these controls use annotated shot changes. The continuous-state reference receives no shot-change input, whereas full Shot3R runs the cut detector and includes the complete integration procedure. The full system is therefore not a single-factor comparison with any intermediate row.

Table 10: **External methods with partial geometry support.** W and WA are in mm; ATE is in m using Sim(3) on EgoBody and SE(3) on EgoHumans. Support is the number of clips with a valid W value, followed by the number with a valid WA value. PromptHMR's ATE on EgoBody covers all 129 clips.

Dataset	Method	Support	W ↓	WA ↓	ATE ↓	IDF1 ↑	Coverage ↑
EgoBody	TRACE	22 / 35	576.4	276.4	–	0.036	0.035
EgoBody	PromptHMR (SPEC)	64 / 87	741.4	336.1	0.302	0.195	0.202
EgoHumans	TRACE	43 / 59	2464.6	749.5	–	0.070	0.072
EgoHumans	PromptHMR (SPEC)	17 / 25	1193.3	487.7	–	0.061	0.062

Table 11: **Three-shot reconstruction on four Harmony4D clips.** N_h is valid human support; IDF1 and Coverage retain all four clips. W/WA are in mm; ATE uses Sim(3) and is in m. Best reported values are bold, including ties.

Method	N_h	W ↓	WA ↓	ATE ↓	IDF1 ↑	Coverage ↑
Human3R	4	2091.9	719.4	**1.618**	0.377	**0.936**
OnlineHMR	3	**1065.7**	**304.3**	3.393	0.055	0.121
Shot3R	4	1304.8	651.2	1.700	**0.434**	**0.936**

Table 12: **Full results for the alignment and registration controls.** (a) Alignment and state propagation on EgoBody. (b) Geometric registration of identical independent-shot predictions. W and WA are in mm; ATE is in m, using Sim(3) on EgoBody and SE(3) on EgoHumans. Bold marks the best value in each column, and the shaded row is full Shot3R.

(a) Spatial alignment and state propagation

Configuration	W ↓	WA ↓	ATE ↓	IDF1 ↑
Continued state	321.1	233.4	0.124	0.870
Reinitialized state	752.6	589.4	1.141	0.834
Alignment + continued state	293.3	227.3	0.082	0.868
Alignment + reinitialized state	352.8	267.7	0.017	0.834
Alignment + reinitialized state + association	349.2	265.2	0.019	0.983
Shot3R	**258.8**	**201.7**	**0.016**	**0.985**

(b) Geometric registration alternatives

	EgoBody			EgoHumans		
Method	W ↓	ATE ↓	IDF1 ↑	W ↓	ATE ↓	IDF1 ↑
Independent shots	752.6	1.141	0.834	1248.1	4.411	0.455
+ pelvis translation	500.2	0.068	0.896	1242.1	3.326	0.506
+ human-joint SE(3)	261.5	0.019	0.896	897.0	1.731	0.510
+ scene FPFH+ICP (offline)	733.2	0.020	0.834	1372.8	3.847	0.455
Shot3R	**258.8**	**0.016**	**0.985**	**866.8**	**1.423**	**0.571**

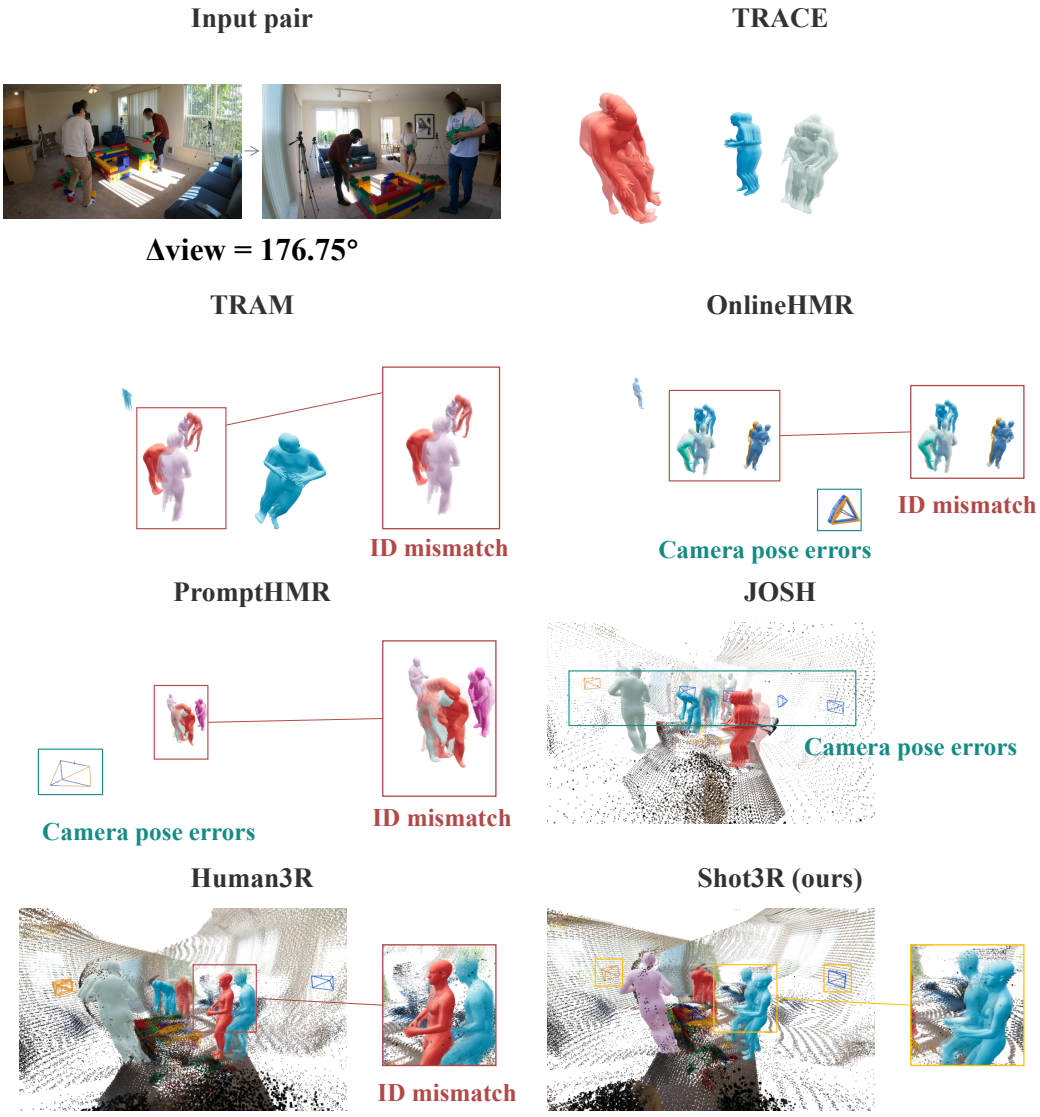

Figure 9: **Full qualitative comparison on EgoBody.** The input pair spans a 176.75° viewpoint change. All seven methods are retained, with human and camera details highlighted by red and teal boxes. The Human3R panel includes the identity-mismatch annotation shown in the main-paper comparison.

Registration alternatives start from identical independent-shot predictions. Pelvis translation and human-joint SE(3) fitting use predictions at the shot change, whereas scene FPFH (Rusu et al., 2009)/RANSAC (Fischler & Bolles, 1981) followed by ICP (Besl & McKay, 1992) accesses complete shots. Failed fits retain the unaligned prediction and remain in evaluation; on EgoHumans, human-joint fitting fails on 16 of 90 clips and scene registration on 53. Shot3R and human-joint fitting have similar W on EgoBody, while Shot3R has higher IDF1 on both datasets.

The pairs behind Table 12 are compared directly. Paired W differences between human-joint fitting and Shot3R are 2.6 mm with a 95% bootstrap interval of $[-40.4, 53.8]$ over 43 recordings and 129 clips on EgoBody, and 24.9 mm with $[-102.0, 134.0]$ over 27 captures and 90 clips on EgoHumans, where positive differences favor Shot3R. Both intervals include zero. Paired means weight

each resampling unit equally, whereas the EgoHumans column is a clip-macro mean, so the paired difference and the difference of the tabulated EgoHumans means need not coincide.

D.2 Camera drift under state propagation

We isolate recurrent-state effects on the retained 90 EgoHumans clips using the same original pre-trained checkpoint and identical post-shot-change inputs, with Shot3R modules disabled. Each shot is recorded by a static camera. For offset k after the first post-shot-change frame b, we evaluate $(C_b^w)^{-1}C_{b+k}^w$ on the world-frame camera poses C_t^w defined in Section 3.3, whose ground-truth value is identity. This relative measure cancels a constant world-frame transform. Table 13 reports translation and rotation averaged over the first K offsets, first within clips and then equally over 27 captures.

Table 13: **State-isolation diagnostic across temporal windows.** C and R denote continued and reinitialized states. Translation is in m and rotation in degrees. The lower error in each continued/reinitialized pair is bold.

Offsets K	Trans. C	Trans. R	Rot. C	Rot. R
10	1.045	**0.019**	14.341	**0.164**
16	1.124	**0.026**	14.578	**0.200**
32	1.256	**0.039**	15.125	**0.272**
49	1.387	**0.047**	15.652	**0.336**

Figure 10 shows the drift curves over all 49 subsequent frames. At the primary window of 16 offsets, reinitialization reduces translation drift by 1.098 m with a 95% capture-bootstrap interval of [0.832, 1.372], and rotation drift by 14.378° with [8.919, 20.494]. All 27 capture means improve for both quantities. This diagnostic supports independent within-shot state evolution. It does not establish the inter-shot world alignment, which is evaluated separately by the alignment controls.

D.3 System-level ablation

Table 14 compares the continuous state, reinitialization alone, camera alignment alone, and the full system under the same frozen protocol and evaluator. Reinitialization and alignment each fall short of the continuous-state baseline, whereas the full system reaches 866.8 mm.

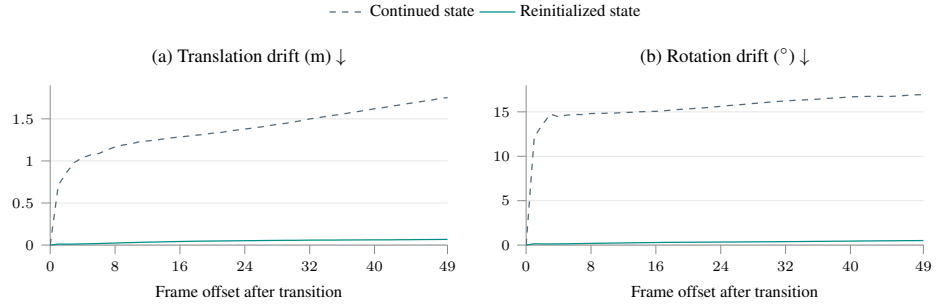

Figure 10: **Effect of recurrent-state propagation on post-shot-change drift.** Both settings use the same checkpoint and post-shot-change images without Shot3R modules. Curves average 27 EgoHumans captures equally and measure poses relative to the first post-shot-change prediction, shown at offset zero. The plot covers all 49 subsequent frames; the main text reports means over the predefined window of offsets 1–16.

Table 14: **System-level ablation on EgoHumans.** All settings share the frozen 90-clip protocol and evaluator, averaged over clips, and all configurations produce a valid W on the same 90 clips. Reinitializing the state alone is worse than the continued-state baseline, and adding camera alignment alone still does not reach it.

Configuration	W ↓
Human3R (continuous state)	1005.1
Reinitialized state	1248.1
Camera alignment	1072.4
Shot3R (full)	**866.8**

E EFFICIENCY

E.1 COMPLETE PREDICTION PIPELINES

Table 15 reports seven methods on five fixed 150-frame Harmony4D streams, comprising 750 input frames. Each required prediction stage runs on one NVIDIA L20. Timing includes fresh process launch, model loading, required preprocessing, reconstruction or optimization, and native output serialization. For multi-stage releases, we sum the required stages. Dataset extraction, metric evaluation, paper-format conversion, and rendering are excluded. All 35 method–clip runs complete. We report mean and population standard deviation in seconds, and compute FPS from the per-clip mean time. Shot3R is causal, and its time covers the complete pipeline.

Table 15: **Complete-pipeline runtime on an NVIDIA L20.** Time is the mean ± population standard deviation over five 150-frame Harmony4D streams, and FPS is 150 divided by the per-clip mean time. Online denotes the evaluated full pipeline, as in Table 1. Rows are grouped by output scope; bold marks the best time and FPS among the pipelines that recover scene geometry.

Method	Online	Time (s) ↓	FPS ↑
Without dense scene geometry			
TRACE	✓	45.0 ± 21.5	3.334
TRAM	×	455.5 ± 60.6	0.329
PromptHMR (SPEC)	×	708.7 ± 209.0	0.212
OnlineHMR	×	1007.8 ± 124.7	0.149
With dense scene geometry			
Human3R	✓	$\mathbf{120.4 \pm 3.2}$	**1.246**
Shot3R	✓	167.8 ± 4.0	0.894
JOSH	×	949.6 ± 286.0	0.158

These systems have different output scopes. JOSH, Human3R, and Shot3R recover scene geometry in addition to people and cameras, whereas the other timed pipelines do not supply comparable dense scene outputs in this evaluation. TRACE provides no evaluated camera trajectory, and PromptHMR provides one on EgoBody and Harmony4D (Tables 10 and 6). Their throughput therefore measures different work.

E.2 EFFECT OF SHOT-CHANGE COUNT

To vary the shot-change count at fixed output length, we repeat a 150-frame EgoBody input once and supply 0, 1, 3, or 5 shot changes on the same 300-frame timeline. Each shot change requires one additional neural evaluation of the incoming frame, giving $300 + C$ evaluations for C shot changes. Each setting uses one warm-up and three measured runs; Table 16 reports median time and the largest peak allocation. Times span 102.6–107.3 s and are not monotonic, so the shot-change count does not measurably change the total time in this range. Peak allocation stays between 5.25 and 5.82 GiB across the tested range.

Table 16: **Shot-change-count control with 300 output frames.** Time is the median of three runs; peak is the maximum allocated memory across those runs. The lowest time and memory and highest FPS are bold.

Shot changes	Neural evaluations	Time (s)	FPS	Peak (GiB)
0	300	105.423	2.846	5.82
1	301	107.286	2.796	5.48
3	303	103.466	2.900	5.30
5	305	**102.585**	**2.924**	**5.25**

F Scope and Limitations

The evaluation concentrates on viewpoint changes within connected environments and on people that remain observable around a shot change. In movies and TV shows, long absences, re-entry, and severe occlusion can break geometric identity association. Because the current matcher lacks long-term appearance-based re-identification and explicit match rejection, a plausible geometric assignment can still be incorrect. Small or weakly visible people in wide shots also remain dependent on the base model's detection and body-estimation quality.

The shot transform expresses people, cameras, and scene points in the shared world frame, but it does not repair missing observations or all errors in the underlying geometry. The primary benchmarks use two-shot clips, and the repeated shot-change check contains only four clips. Longer streams, gradual edits, and unrelated scene changes remain outside the present validation. The additional results also retain the observed tradeoffs between Coverage, human error, and camera error.

G Qualitative 4D human-scene reconstruction results

G.1 Outdoor sequence

Figure 11 complements the film-footage example in Figure 5 with an outdoor single-person sequence from a public instructional video.

G.2 Reconstruction examples across datasets

Figures 12–14 show examples from AIST++, Harmony4D, and EgoBody. Each example overlays four selected instants and places their input frames below, with two frames on either side of the shot change.

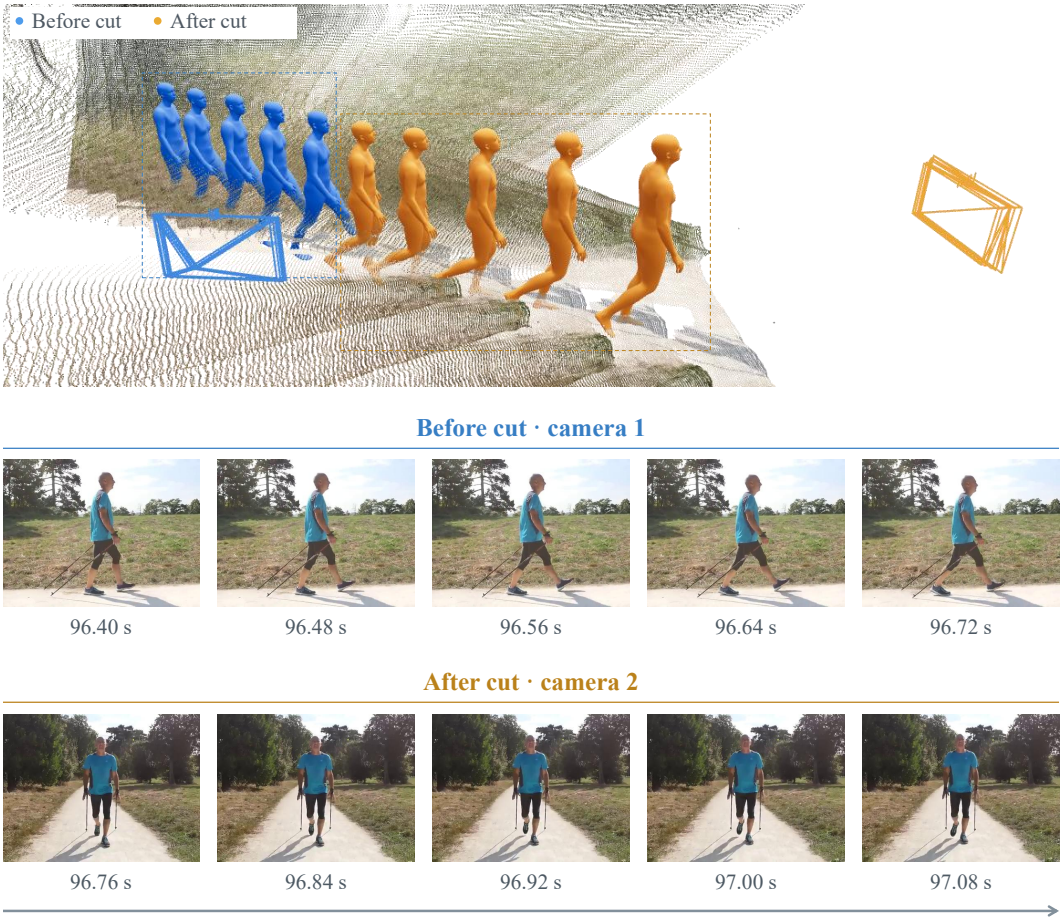

Figure 11: **Qualitative 4D human-scene reconstruction results on an outdoor sequence.** The upper view shows the scene, human motion, and cameras in a shared world frame. Blue and gold distinguish the two shots of the same person. Ten input frames run chronologically from left to right in each row, with the pre-cut row above the post-cut row. The shot change occurs at 96.76 s.

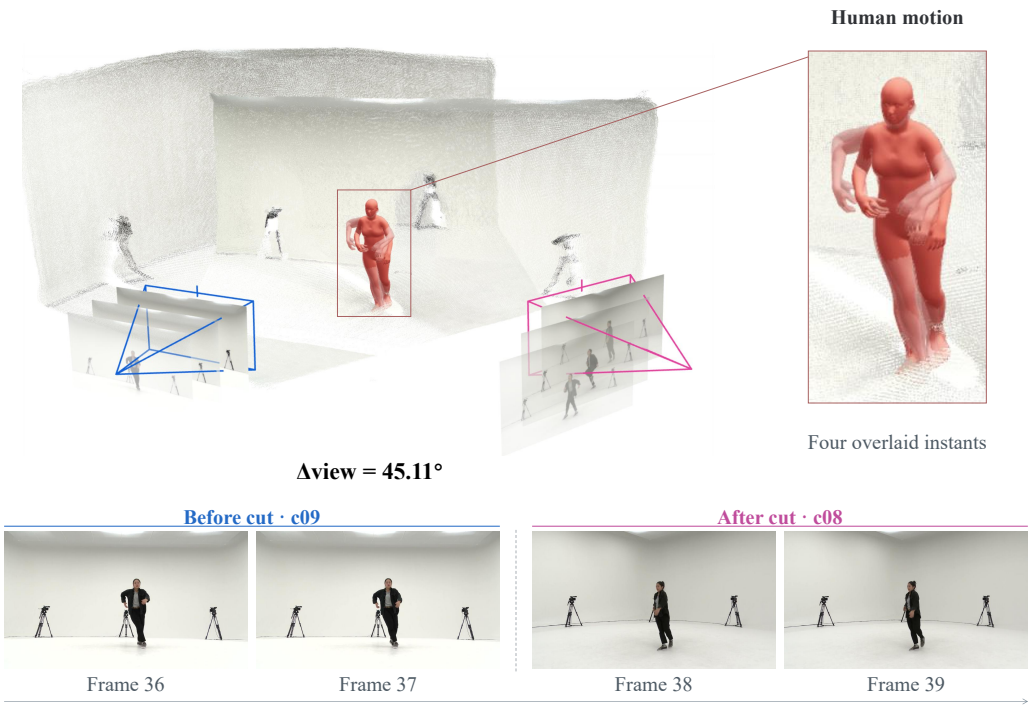

Figure 12: **Qualitative 4D human-scene reconstruction results on AIST++.** This example uses a prediction-only single-person torso anchor that applies one unit-scale $\mathrm{SE}(3)$ transform to the entire post-cut shot, without ground-truth alignment. The camera changes from c09 to c08 ($45.11°$); the inputs are source frames 36–39. Blue and magenta distinguish the pre- and post-cut camera groups. The connected inset enlarges the four overlaid human poses.

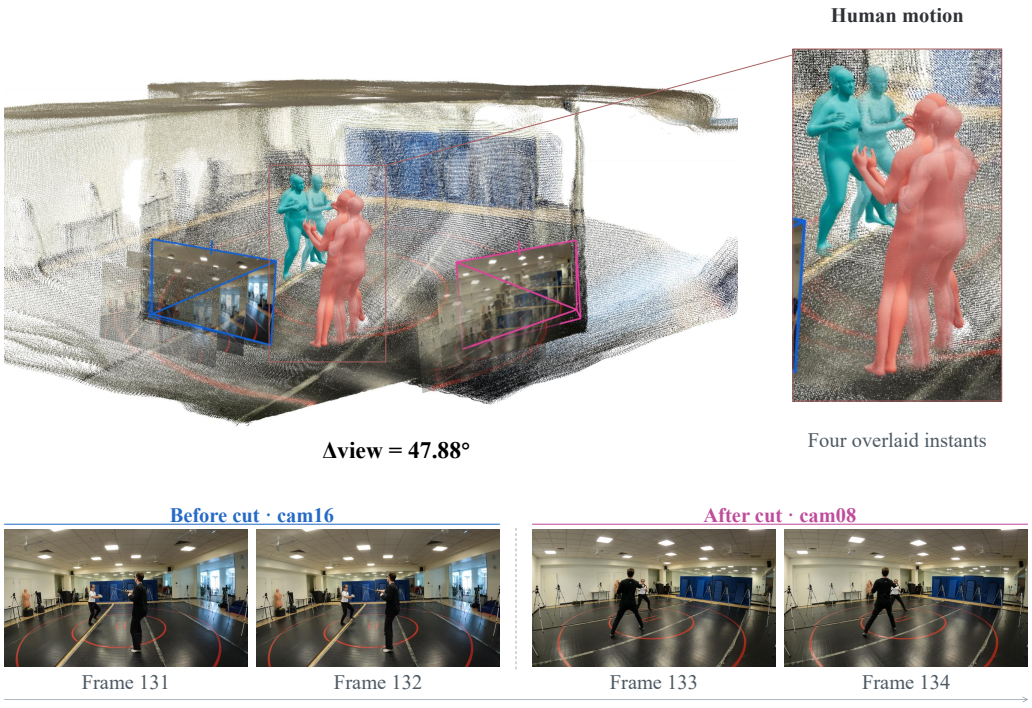

Figure 13: **Qualitative 4D human-scene reconstruction results on Harmony4D.** Standard multi-person Shot3R is shown on a two-person sword interaction with a cam16-to-cam08 shot change (47.88°). Source frames 131–134 appear below the reconstruction. The human poses are overlaid with increasing opacity and enlarged in the connected inset; blue and magenta distinguish the pre- and post-cut camera groups.

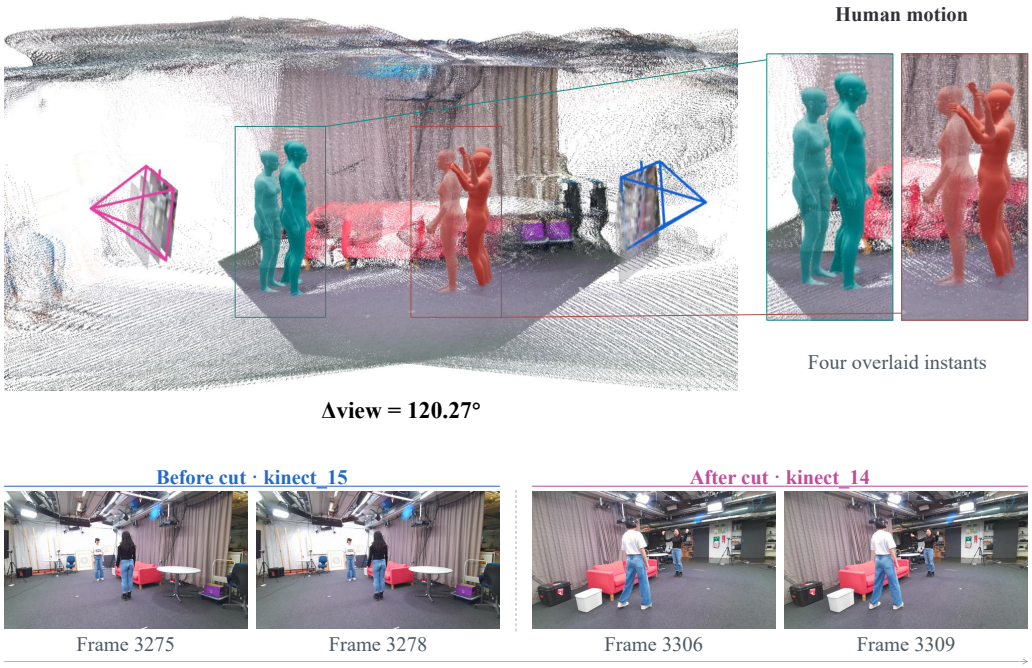

Figure 14: **Qualitative 4D human-scene reconstruction results on EgoBody.** Standard multi-person Shot3R is shown on a two-person indoor sequence with a kinect_15-to-kinect_14 shot change (120.27°). Separate insets enlarge the two people's overlaid poses. The inputs are source frames 3275, 3278, 3306, and 3309, ordered from left to right. Blue and magenta distinguish the pre- and post-cut camera groups.