CAHI4D: CROSS-AGE HUMAN INTERACTION 4D RECONSTRUCTION FROM MONOCULAR VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Age-dependent morphology exacerbates scale–depth ambiguity in monocular multi-person reconstruction, while adult-centric priors and occlusion of interaction cues further hinder cross-age interaction recovery. We present Cross-Age Human Interaction 4D Reconstruction (CAHI4D), which uses a reliable reference and validated interactions for geometric anchoring before fitting age-dependent shape and refining interaction motion. A vision–language model (VLM) proposes support and contact hypotheses from visual context; geometric and temporal checks then determine which relations can constrain reconstruction. Depth- and Interaction-Guided Geometric Anchoring (DIGA) uses a reliable scene-supported reference to align the shared scene scale, and then places a geometrically underconstrained participant through validated interactions and corrects body scale from the induced projection change. With fixed placement and overall body scale, shape fitting adapts age-dependent proportions. On the fixed geometry, Kinematic-Topology-Guided Interaction Correction (KTIC) maps directed surface corrections to participant-specific kinematic chains and reconciles corrections sharing joint degrees of freedom; contact and temporal terms constrain the same motion fit. CAHI4D achieves the lowest errors among the compared methods on Hi4D, 3DPW, and CMU toddler. In particular, on CMU, Root MPJPE and Joint-PA MPJPE fall by 21.0% and 49.1%, respectively. Code will be released.

1 INTRODUCTION

Recent monocular methods recover articulated human bodies and global trajectories from video (Ye et al., 2023; Wang et al., 2024; Wen et al., 2025; Kang et al., 2026), but plausible per-person reconstructions alone do not guarantee a coherent 4D interaction. As illustrated in Figure 1, participants in cross-age scenes differ in body proportions while their relative placement, motion, and contact must agree in one 3D space throughout the sequence. Close interactions are especially demanding because an error in one participant's scale or depth can make the pair's relative placement and contact inconsistent. Cross-age human interaction 4D reconstruction must therefore recover both individual geometry and the evolving relationship between participants.

Age-dependent morphology makes this relative geometry difficult to infer. Under perspective projection, a small image extent can indicate a shorter body, greater depth, or both; body scale and viewing-direction placement can therefore compensate for each other. Adult-centric body priors can also explain a child's different proportions through erroneous scale or pose. Age-inclusive body models expand shape coverage (Hesse et al., 2020; Brégier et al., 2026), while image-based estimators address cross-age shape and relative depth (Sun et al., 2022; Bravo-Sánchez et al., 2026) and dynamic multi-person methods recover global trajectories (Ye et al., 2023; Wen et al., 2025; Kang et al., 2026). Yet these advances alone do not ensure sequence-consistent relative geometry during cross-age interactions. As Figure 1 shows, per-person depth can be unreliable. In our method, the scene-supported adult provides a reliable depth reference, while validated interaction cues constrain the child's placement relative to that adult over time.

The body regions underlying these interaction cues are often occluded. Scene cues have improved human–scene reconstruction (Liu et al., 2026), while contact and semantic cues have improved interpersonal reconstruction (Fieraru et al., 2020; Müller et al., 2024; Xia et al., 2026). Yet visible motion and the overall interaction can imply support or contact beyond what a single frame reveals.

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.

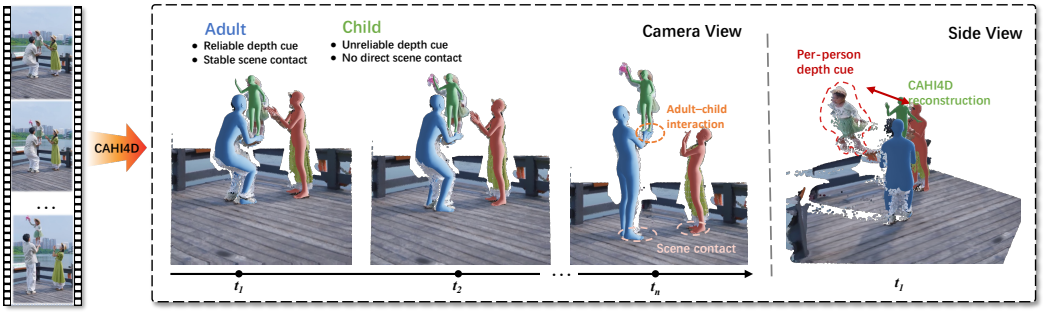

Figure 1: **Cross-age human interaction 4D reconstruction from monocular video.** The scene-supported adult serves as the reference for interaction-guided child placement, and the side view at t_1 contrasts a per-person depth cue with the CAHI4D reconstruction.

When an adult visibly supports a child's upper body, for example, the carrying pose may suggest an occluded forearm-to-thigh support. Such a relation is a hypothesis, not a measured contact point; surface geometry and temporal consistency must determine whether it can constrain reconstruction.

We present **Cross-Age Human Interaction 4D Reconstruction (CAHI4D)**, based on the insight that validated interactions turn a better-observed participant's evidence into relative-position constraints for an underconstrained participant; the projection change induced by placement refinement then constrains body scale before shape fitting and interaction-motion refinement. For this purpose, a vision-language model (VLM) proposes interpersonal support and contact hypotheses from visual cues and interaction context, and geometric and temporal checks admit only supported relations. As shown in Figure 2, *Depth- and Interaction-Guided Geometric Anchoring* (DIGA) aligns monocular depth to a reliable scene-supported reference and uses validated interactions to place underconstrained participants. Because moving the participant in depth changes its projected size, DIGA uses the before-and-after projected-size ratio to correct body scale across the sequence. These stages first anchor relative placement and overall scale, then fit participant-specific proportions at fixed placement and overall body scale, and finally refine residual interaction motion on the fitted geometry.

The geometric anchor constrains overall body scale while leaving age-dependent proportions free to vary. Anny–SOMA represents those proportions on a common topology (Brégier et al., 2026; Saito et al., 2026), allowing an underconstrained participant's own image and body-attribute evidence to refine shape at fixed placement and overall scale. Fitting these proportions before motion refinement lets contact and penetration constraints act on participant-specific surfaces, reducing the need for motion residuals to absorb shape error. The resulting surface corrections can demand incompatible rotations at the same joint. Multi-task inverse kinematics provides a precedent for reconciling competing kinematic constraints (Baerlocher & Boulic, 2004). During interaction-motion fitting on the fixed geometric base, *Kinematic-Topology-Guided Interaction Correction* (KTIC) maps these directed corrections to participant-specific kinematic chains and jointly reconciles corrections on overlapping chains into one rotation target per affected joint; contact and temporal terms constrain the same motion optimization. CAHI4D achieves the lowest error among the compared methods across the reported metrics on Hi4D, 3DPW, and CMU toddler; on CMU, Root MPJPE and Joint-PA MPJPE decrease by 21.0% and 49.1%, respectively, relative to the strongest evaluated baseline.

Our contributions are:

- We introduce CAHI4D for monocular cross-age human interaction 4D reconstruction, using validated VLM-proposed interaction hypotheses, including occluded support and contact, to establish shared geometry before fitting age-dependent shape and refining interaction motion.
- We propose DIGA, which places underconstrained participants relative to reliable references using aligned depth and validated interactions, then corrects scale from the induced projection change to establish sequence-consistent geometric anchoring.
- We propose KTIC, which maps directed surface corrections to participant-specific kinematic chains and jointly coordinates corrections affecting the same participant-indexed joint variables while preserving the anchored geometric base.

- Experiments on Hi4D, 3DPW, and CMU toddler show the lowest errors among the compared methods, with the largest relative gains on CMU toddler.

2 Related Work

Cross-age human modeling and reconstruction. SMPL and SMPL-X provide parametric body spaces used in monocular reconstruction (Loper et al., 2015; Pavlakos et al., 2019). SMIL models infant shape (Hesse et al., 2020); AGORA extends SMPL-X to fit children using an infant–adult template interpolation (Patel et al., 2021); BEV uses SMPL+A for infant-to-adult body variation (Sun et al., 2022). Anny spans the lifespan (Brégier et al., 2026), and SOMA provides a unified topology and articulation space (Saito et al., 2026). BEV addresses body-size/depth ambiguity across ages (Sun et al., 2022), while AionHMR targets child recovery (Chatzichristodoulou et al., 2026); Anny-Fit additionally fits multiple people in an image using metric depth, image evidence, and vision–language-derived body attributes (Bravo-Sánchez et al., 2026). Other studies examine scale, shape–position estimation, and uncertainty (Ugrinovic et al., 2021; Wang et al., 2025a; Brégier et al., 2025; Song et al., 2026; Sengupta et al., 2021; Zhang et al., 2023). HARMONI analyzes caregiver–child interactions in video, but focuses on behavioral analysis (Weng et al., 2025). These advances improve age coverage or individual recovery, while sequence-consistent relative geometry in cross-age interactions remains less explored.

Multi-person reconstruction and close interaction. Single-image methods recover multiple people (Baradel et al., 2024), while video methods reconstruct and track multiple people through time (Goel et al., 2023; Fiche et al., 2026; Yuan et al., 2022; Ye et al., 2023; Chen et al., 2026). Complementary video methods estimate global motion for an individual (Wang et al., 2024; Shin et al., 2024; Shen et al., 2024). Contact-rich datasets and learned priors model interpersonal configurations (Fieraru et al., 2020; 2025; Yin et al., 2023; Müller et al., 2024). Close-interaction methods use proxemic, appearance, or physical constraints to improve relative placement and contact (Huang et al., 2024; Ugrinovic et al., 2024; Huang et al., 2025). Newer approaches add group-aware geometry or interaction-conditioned recovery (Kim et al., 2026; Zheng et al., 2026; Xia et al., 2026). Language and vision–language models provide pose priors, contact reasoning, reconstruction feedback, or body-shape prompts (Subramanian et al., 2025; Dwivedi et al., 2025; Xu et al., 2025; Wang et al., 2025b). Multi-task inverse kinematics reconciles competing joint constraints (Baerlocher & Boulic, 2004). CAHI4D addresses the coupled cross-age problem by anchoring relative geometry across a sequence, fitting shape under that anchor, and then correcting interaction-motion residuals.

3 Method

3.1 Overview and Human Representation

Given monocular frames $\{I_t\}_{t=1}^{T}$ containing P interacting participants, Cross-Age Human Interaction 4D Reconstruction (CAHI4D) reconstructs sequence-consistent body shapes and motions in a shared 3D space. Anny represents age-dependent stature and proportions, while SOMA provides common topology and kinematics for inter-person surface correspondences and motion correction (Brégier et al., 2026; Saito et al., 2026). For participant p, $\boldsymbol{\phi}_p$ denotes the sequence-shared Anny phenotype controls and $\boldsymbol{\theta}_{p,t}$ the frame-dependent non-root local pose. The pelvis-centered Anny–SOMA mesh $\overline{\mathbf{Y}}_{p,t}(\boldsymbol{\phi}_p, \boldsymbol{\theta}_{p,t})$ excludes root rotation, so its shared-space surface is

$$\mathbf{X}_{p,t} = \mathbf{r}_{p,t} + \mathbf{R}_{p,t}\overline{\mathbf{Y}}_{p,t}(\boldsymbol{\phi}_p, \boldsymbol{\theta}_{p,t}), \tag{1}$$

where $\mathbf{r}_{p,t}$ is the pelvis position and $\mathbf{R}_{p,t} \in SO(3)$ applies root rotation once in the shared coordinate system.

Fixed observations comprise ViTPose keypoints, SAM 2 masks, VGGT-Ω depth and camera/scene cues, and GeoCalib gravity (Xu et al., 2022; Ravi et al., 2024; Wang et al., 2026; Veicht et al., 2024). VIMO (Wang et al., 2024) supplies frame-wise SMPL estimates: aggregated canonical shape initializes a sequence-shared Anny phenotype, root-normalized surfaces initialize SOMA poses, and frozen camera extrinsics map roots into the shared space.

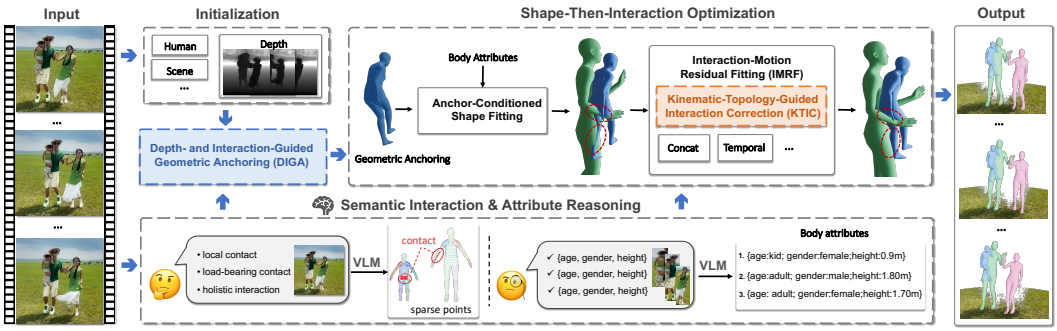

Figure 2: **CAHI4D overview.** Initialization extracts human, scene, and depth cues from monocular video, and VLM reasoning proposes interaction hypotheses and estimates body attributes. Geometric anchoring is followed by shape fitting and interaction-motion residual fitting to produce 4D reconstruction.

Semantic Interaction and Attribute Reasoning. We select candidate keyframes when the participants' 2D poses are close, then pass them in temporal order with consistent identities to the vision–language model (VLM) Qwen3-VL (Bai et al., 2025). It assesses sequence-level interaction, estimates age and scale-invariant body attributes, and proposes per-joint interaction states and cross-person contact/support hypotheses, including occluded support. Attributes guide shape fitting, while interaction hypotheses guide geometric anchoring and motion fitting only after geometric and temporal validation. The VLM supplies neither metric measurements nor exact contact points; Appendix A.1 gives the prompt.

The pipeline proceeds through Depth- and Interaction-Guided Geometric Anchoring (DIGA), shape fitting, and interaction-motion fitting. We separate geometric placement from a later root residual and initialized local pose from its rotation residual:

$$\mathbf{r}_{p,t} = \mathbf{r}^{\text{geo}}_{p,t} + \Delta\mathbf{r}^{\text{mot}}_{p,t}, \qquad \boldsymbol{\theta}_{p,t} = \boldsymbol{\theta}^{0}_{p,t} \oplus \Delta\boldsymbol{\theta}^{\text{mot}}_{p,t}, \tag{2}$$

where $\Delta\mathbf{r}^{\text{mot}}$ is in the shared coordinate system and $\oplus$ denotes a local right-multiplicative rotation update. DIGA estimates placement and a projection-consistent auxiliary scale, and shape fitting absorbs the anchor-derived stature into the sequence-shared phenotype, so the auxiliary scale is not reapplied to the final mesh. With the base geometry fixed, motion fitting optimizes the residuals, while Kinematic-Topology-Guided Interaction Correction (KTIC) coordinates surface-driven rotation targets.

3.2 Depth- and Interaction-Guided Geometric Anchoring

Monocular depth couples relative placement and body size, so even a fully visible participant's absolute size cannot be determined independently. DIGA therefore combines complementary cues with distinct roles: image and mask evidence establish reliable visibility, scene evidence supplies support, depth constrains shared-space placement, geometrically validated interpersonal support links a geometrically underconstrained participant to a reliable reference, and body-attribute cues restrict the feasible phenotype domain for subsequent anchor-conditioned shape fitting. None of these non-metric cues independently supplies absolute scale. During placement fitting, DIGA holds the initialized phenotype $\boldsymbol{\phi}^0$ and auxiliary ratio $s^0 = 1$ fixed, then uses the resulting projection change to estimate the sequence-level anchoring ratio $s^\star$. Let $\mathcal{D}_{\text{A}}$ be the valid domain of the selected Anny phenotype controls and $H_{\text{A}}(\boldsymbol{\phi}) > 0$ the canonical stature of the resulting mesh. Figure 3 (left) illustrates the DIGA stages.

Reference Selection and Depth Alignment. In each component formed by VLM-proposed support candidates, DIGA selects the participant with the most reliable mask, depth, and scene evidence as a candidate reference. In Figure 1, this is the scene-supported adult with well-observed joints and a reliable depth cue. We first align scene scale to this reference in the current shared gauge, and then align monocular depth to the reference depth. After reference fitting, VLM-proposed relations become placement constraints only if surface-distance, normal, overlap, aligned-depth-order, and temporal checks yield stable correspondences. We write $i \to j$ when reference i supplies such an

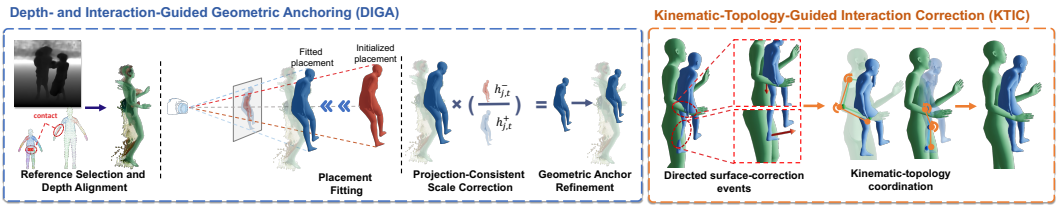

Figure 3: Detailed stages of Depth- and Interaction-Guided Geometric Anchoring (DIGA, left) and Kinematic-Topology-Guided Interaction Correction (KTIC, right).

admitted interaction-derived constraint for target j. This interaction constrains the child relative to the reference, while other participants use their own image and depth cues. Appendix A.2 gives the gates and residuals.

Placement Fitting. Let $\boldsymbol{\phi}_j^0$, $\boldsymbol{\theta}_{j,t}^0$, $s_j^0 = 1$, and $\bar{\mathbf{r}}_{j,t}$ denote target j's initialized phenotype, local pose, auxiliary anchoring ratio, and shared-space root position. For a frame with a nonempty admitted correspondence set $\mathcal{K}_{j,t}$, $k \in \mathcal{K}_{j,t}$ indexes a valid body-part surface correspondence and i_k identifies its reference participant. Let $\mathbf{a}_{i_k,k,t}$ be the detached reference surface anchor and $\bar{\mathbf{q}}_{j,k,t}(\boldsymbol{\phi}_j)$ the corresponding pelvis-centered point on the target's Anny–SOMA mesh at the fixed pose $\boldsymbol{\theta}_{j,t}^0$. Their support-implied root under a temporary ratio s is

$$\widehat{\mathbf{r}}_{j,t}^{\mathrm{sup}}(\boldsymbol{\phi}_j, s) = \sum_{k \in \mathcal{K}_{j,t}} \bar{w}_{k,t} \left[\mathbf{a}_{i_k,k,t} - \mathbf{R}_{j,t} s \bar{\mathbf{q}}_{j,k,t}(\boldsymbol{\phi}_j) \right], \qquad \bar{w}_{k,t} \geq 0, \quad \sum_{k \in \mathcal{K}_{j,t}} \bar{w}_{k,t} = 1, \tag{3}$$

where the normalized $\bar{w}_{k,t}$ yields one confidence-weighted root proposal. Holding the reference, $\mathbf{R}$, $\boldsymbol{\theta}^0$, camera, scene, $\boldsymbol{\phi}_j^0$, and $s_j^0 = 1$ fixed, DIGA fits $\mathbf{r}_{j,1:T}^+$ from image, aligned-depth, extent, scene, support, and temporal evidence. Aligned depth constrains placement by comparing rendered body depth with the frozen observation when calibrated, or by enforcing the observed depth ordering without metric calibration. Appendix A.2 gives the residuals.

Projection-Consistent Scale Correction. Position fitting changes depth and hence projected size, so DIGA compensates for this change to preserve the original image extent. Crucially, size is measured only on the target's isolated amodal silhouette: scene and other participants are excluded, while both renders share the original uncropped canvas, camera, $\boldsymbol{\phi}_j^0$, $\boldsymbol{\theta}_{j,t}^0$, $\mathbf{R}_{j,t}$, and $s_j^0 = 1$. Let $\widehat{M}_{j,t}^{\mathrm{iso}}(\mathbf{u}; \mathbf{r}, s)$ render the temporary surface $\mathbf{r} + \mathbf{R}_{j,t}[s\overline{\mathbf{Y}}_{j,t}(\boldsymbol{\phi}_j^0, \boldsymbol{\theta}_{j,t}^0)]$. The projected image length and its two snapshots are

$$\begin{aligned} \mathcal{H}_{j,t}(\mathbf{r}, s) &= \sqrt{\sum_{\mathbf{u}} \widehat{M}_{j,t}^{\mathrm{iso}}(\mathbf{u}; \mathbf{r}, s)}, \\ h_{j,t}^- = \mathcal{H}_{j,t}(\bar{\mathbf{r}}_{j,t}, s_j^0), &\qquad h_{j,t}^+ = \mathcal{H}_{j,t}(\mathbf{r}_{j,t}^+, s_j^0). \end{aligned} \tag{4}$$

Taking the square root converts area to image length. Under weak perspective, its ratio gives the required linear correction; under full perspective, it is a first-order initializer for the bounded refinement below:

$$\chi_{j,t} = \frac{h_{j,t}^- + \varepsilon}{h_{j,t}^+ + \varepsilon}, \tag{5}$$

Here, $\varepsilon > 0$ stabilizes the ratio. Moving further yields $\chi_{j,t} > 1$ and a larger anchoring ratio. Over valid frames $\mathcal{T}_j$, the sequence-shared log-ratio correction $\widehat{\eta}_j$ robustly pools $\log \chi_{j,t}$ using normalized image–support–projected-extent weights $w_{j,t}$ and penalty ρ:

$$\widehat{\eta}_j = \operatorname*{argmin}_{\eta} \sum_{t \in \mathcal{T}_j} w_{j,t} \rho(\eta - \log \chi_{j,t}), \qquad \widetilde{s}_j = s_j^0 \exp(\widehat{\eta}_j), \tag{6}$$

where $\sum_{t \in \mathcal{T}_j} w_{j,t} = 1$. The resulting $\widetilde{s}_j$ is a dimensionless initializer and adds no metric scale.

Geometric Anchor Refinement. Because scaling about the pelvis moves a supported surface point, a bounded one-dimensional refinement initialized by $\widetilde{s}_j$ recomputes its root while preserving the accepted placement offset:

$$\mathbf{r}^{\mathrm{geo}}_{j,t}(s) = \widehat{\mathbf{r}}^{\mathrm{sup}}_{j,t}(\boldsymbol{\phi}^0_j, s) + \mathrm{sg}\big[\mathbf{r}^{+}_{j,t} - \widehat{\mathbf{r}}^{\mathrm{sup}}_{j,t}(\boldsymbol{\phi}^0_j, s^0_j)\big]\,. \tag{7}$$

Here sg stops gradients through the accepted offset. Full-perspective refinement accepts $s^\star_j$ only if image consistency improves, support remains valid, and $s^\star_j H_{\mathrm{A}}(\boldsymbol{\phi}^0_j)$ is attainable in $\mathcal{D}_{\mathrm{A}}$; otherwise, it backtracks. On supported frames, $\mathbf{r}^{\mathrm{geo}\star}_{j,t} = \mathbf{r}^{\mathrm{geo}}_{j,t}(s^\star_j)$, with smooth extension elsewhere. The accepted ratio defines

$$\overline{\mathbf{Y}}^{\mathrm{anc}}_{j,t} = s^\star_j \overline{\mathbf{Y}}_{j,t}(\boldsymbol{\phi}^0_j, \boldsymbol{\theta}^0_{j,t}). \tag{8}$$

Because $\boldsymbol{\phi}^0_j$ is fixed, this retains the initialized proportions: for example, an initial adult proxy for a child yields a scaled adult anchor, not a recovered child. The triplet $(\mathbf{r}^{\mathrm{geo}\star}_{j,1:T}, \mathbf{R}_{j,1:T}, \overline{\mathbf{Y}}^{\mathrm{anc}}_{j,1:T})$ is passed to shape fitting. Appendix A.2 gives solver details and fallbacks.

3.3 Shape-Then-Interaction Optimization

Anchor-Conditioned Shape Fitting (ACSF). With $\mathbf{r}^{\mathrm{geo}\star}_{j,1:T}$, $\mathbf{R}_{j,1:T}$, and $\boldsymbol{\theta}^0_{j,1:T}$ fixed, shape fitting updates only the sequence-shared phenotype:

$$\boldsymbol{\phi}^\star_j = \operatorname*{argmin}_{\boldsymbol{\phi}_j \in \mathcal{D}_{\mathrm{A}}} \mathcal{L}^{\mathrm{shape}}_{\mathrm{evid},j}(\boldsymbol{\phi}_j) \quad \text{s.t.} \quad H_{\mathrm{A}}(\boldsymbol{\phi}_j) = s^\star_j H_{\mathrm{A}}(\boldsymbol{\phi}^0_j). \tag{9}$$

Here, $\mathcal{L}^{\mathrm{shape}}_{\mathrm{evid},j}$ evaluates image, depth, attribute, and support evidence on the actual candidate mesh at fixed placement. The constraint preserves anchor stature while allowing age-dependent proportions to change, so subsequent contact and penetration constraints act on the fitted body, reducing the need for motion residuals to absorb shape error. Eq. 1 then uses $\overline{\mathbf{Y}}_{j,t}(\boldsymbol{\phi}^\star_j, \boldsymbol{\theta}_{j,t})$ with Eq. 2, and $s^\star_j$ is not applied again. Appendix A.5 gives the residuals and schedule.

Interaction-Motion Residual Fitting (IMRF). After shape fitting, $\mathbf{r}^{\mathrm{geo}\star}$, $\mathbf{R}$, $\boldsymbol{\phi}^\star$, and $\boldsymbol{\theta}^0$ are frozen. The motion stage optimizes the residuals in Eq. 2, permitting bounded horizontal root residuals only for non-reference participants selected by the root-update mask. KTIC constructs coordinated rotation targets from directed surface corrections, while image evidence, penetration, persistent contact, root trust, and temporal regularization constrain the same residual solve. These are simultaneous terms in Eq. 28, not separate optimization stages.

3.4 Kinematic-Topology-Guided Interaction Correction

KTIC resolves incompatible surface corrections whose kinematic chains share joint degrees of freedom by coordinating them into one rotation target per affected joint. Figure 3 (right) illustrates this process.

Directed surface-correction events. Validated dependencies propose material contact, while visible overlap triggers penetration tests; geometric and temporal checks retain directed events $e = (o_e \leftarrow q_e, \mathcal{P}_e, \mathcal{I}_e, \mathcal{C}_e)$. Here, o_e owns the correction, q_e is the detached target, and $(\mathcal{P}_e, \mathcal{I}_e, \mathcal{C}_e)$ specify its patch, interval, and participant-indexed chain. This ownership arrow differs from a DIGA support edge. Each retained violation yields a clipped displacement $\mathbf{b}_{e,t}$ and surface Jacobian $\mathbf{J}_{e,t}$; Appendix A.4.1 gives their construction.

Kinematic-topology coordination. Let $\mathcal{E}_t$ be the active events at frame t, with two events adjacent when their participant-indexed chains share a joint degree of freedom. For each component $\mathcal{Q}_t \subseteq \mathcal{E}_t$, $\boldsymbol{\xi}_{\mathcal{Q}_t,t}$ contains every owner-joint increment once and $\mathbf{S}_e$ selects event e's chain. Let $\ell^\star_{e,t} > 0$ be the frozen fitted-body scale for event e. With confidence $g_{e,t} \geq 0$, anisotropic weight $\mathbf{W}_{e,t} \succeq 0$, dimensionless damping $\mu_{\mathrm{DLS}} > 0$, and $\|\mathbf{z}\|^2_{\mathbf{W}} = \mathbf{z}^\top \mathbf{W} \mathbf{z}$, KTIC solves

$$\boldsymbol{\xi}^\star_{\mathcal{Q}_t,t} = \operatorname*{argmin}_{\boldsymbol{\xi}} \sum_{e \in \mathcal{Q}_t} g_{e,t} \left\| \frac{\mathbf{J}_{e,t}\mathbf{S}_e\boldsymbol{\xi} - \mathbf{b}_{e,t}}{\ell^\star_{e,t}} \right\|^2_{\mathbf{W}_{e,t}} + \mu_{\mathrm{DLS}} \|\boldsymbol{\xi}\|^2_2. \tag{10}$$

Participant indices keep same-named joints distinct. Backtracking and the $SO(3)$ exponential map produce one detached target $\widehat{\mathbf{Q}}_{p,a,t}$ and confidence $\bar{g}_{p,a,t}$ per affected joint; Appendix A.4.1 gives the mapping.

Rotation Targets in IMRF. The KTIC targets $\widehat{\mathbf{Q}}_{p,a,t}$ are detached within each motion-update block and guide the motion residuals through a weighted rotation loss. For each joint a, the current rotation $\mathbf{Q}_{p,a,t} \in SO(3)$ is $\mathbf{Q}^0_{p,a,t}\,\mathrm{Exp}([\Delta\boldsymbol{\theta}^{\mathrm{mot}}_{p,a,t}]_\times)$ and is not an independent variable. We use the principal $\mathrm{Log}(\cdot)^\vee$, norm-clip right-multiplicative target increments, and reject any inner motion step whose target–current angle reaches $\pi - 10^{-4}$. The target–current discrepancy is

$$\mathcal{L}_{\mathrm{KTIC}} = \sum_{p,a,t:\,\bar{g}_{p,a,t}>0} \bar{g}_{p,a,t} \| \mathrm{Log}(\mathrm{sg}(\widehat{\mathbf{Q}}_{p,a,t})^\top \mathbf{Q}_{p,a,t})^\vee \|_2^2. \tag{11}$$

This term is optimized jointly with image, penetration, contact, root, and temporal terms in Eq. 28; Appendix A.5 gives the schedule.

4 Experiments

Table 1: Quantitative comparison and cumulative CAHI4D framework ablation on Hi4D, 3DPW, and CMU toddler.

Method	Body model	Temp.	Hi4D				3DPW				CMU	
			MPJPE↓	PA-MPJPE↓	MPVPE↓	Inter.↓	MPJPE↓	PA-MPJPE↓	MPVPE↓	Inter↓	Root MPJPE↓	Joint-PA MPJPE↓
BEV (CVPR'22)	SMPL+A	×	85.2	60.5	106.2	85.7	82.6	59.2	99.6	203.6	113.9	161.7
Anny-Fit (CVPR Findings'26)	Anny	×	109.1	55.6	120.1	92.9	130.9	69.1	150.7	412.7	100.2	156.3
Multi-HMR 2 (ECCV'26)	Anny	×	69.7	47.2	83.2	78.8	70.1	45.9	82.7	162.2	102.3	225.5
CloseApp (CVPR'25)	SMPL	✓	59.1	44.3	72.0	80.2	64.5	45.6	75.2	96.4	150.0	299.6
Human3R (ICLR'26)	SMPL-X	✓	97.3	61.4	122.6	125.1	95.4	53.9	112.2	479.4	123.9	173.7
JOSH (ICLR'26)	SMPL	✓	57.5	40.8	72.9	72.6	64.7	40.9	76.0	284.0	141.9	261.2
Cumulative CAHI4D framework												
(1) VIMO Init.	SMPL	✓	57.7	41.6	72.6	106.9	66.2	41.5	78.0	232.2	141.2	276.9
(2) + DIGA	Anny–SOMA	✓	56.5	41.5	71.1	73.3	63.2	39.5	76.7	107.9	81.5	83.2
(3) + ACSF	Anny–SOMA	✓	56.3	41.4	70.8	74.2	63.1	40.5	75.6	107.1	79.3	79.8
(4) + IMRF (CAHI4D)	Anny–SOMA	✓	**55.1**	**40.6**	**69.3**	**71.1**	**63.0**	**39.2**	**74.6**	**93.1**	**79.2**	**79.6**



Figure 4: **Qualitative comparison with existing methods.** Rows show Hi4D adult–adult, CMU adult–child, and in-the-wild adult–child interactions. Each method is shown as an image overlay and from additional viewpoints.

4.1 Experimental Setup

Datasets and metrics. We evaluate CAHI4D on Hi4D (Yin et al., 2023), 3DPW (von Marcard et al., 2018), and CMU Panoptic toddler sequences (Joo et al., 2015; Bravo-Sánchez et al., 2026). For Hi4D and 3DPW, we follow CloseApp's splits and protocol (Huang et al., 2025) and report MPJPE, PA-MPJPE, MPVPE, and Inter (Huang et al., 2024), measuring joint-position, Procrustes-aligned

joint, mesh-vertex, and cross-person joint-distance errors, respectively. For CMU, we evaluate 13 selected monocular adult–child clips (2,078 frames) from five sequences. As CMU provides 3D joints but no ground-truth meshes, we follow Anny-Fit (Bravo-Sánchez et al., 2026) and report Root MPJPE (joint error after aligning each participant's root separately) and Joint-PA MPJPE (joint error after one shared Procrustes alignment of both participants, retaining their relative placement). All compared methods use the same selected views and frames; Appendix A.6 details their selection. Lower is better for all metrics.

Penetration-depth error. For the Hi4D ablation, $\mathbf{d}^{\mathrm{pred}}$ and $\mathbf{d}^{\mathrm{GT}}$ collect vertex penetration depths in millimeters, each computed against the other person's mesh within the prediction or ground truth, respectively, with zero for non-penetrating vertices. We report $\mathrm{PD\text{-}MAE} = \mathrm{MAE}(\mathbf{d}^{\mathrm{pred}}, \mathbf{d}^{\mathrm{GT}})$ in millimeters over corresponding vertices of both people and all evaluated frames.

Implementation details. CAHI4D processes each video independently, starting from VIMO estimates mapped to Anny–SOMA. All CAHI4D experiments were run on a single NVIDIA GeForce RTX 4090 D GPU. We use stage-specific optimization with fixed settings across the evaluated sequences; schedules, loss weights, and ablation controls are given in Appendix A.5.

4.2 Comparison with Existing Methods

Quantitative analysis. As shown in Table 1, CAHI4D obtains the lowest error in every reported metric among the compared methods. Relative to the strongest baseline in each column, it reduces the four Hi4D errors by 2.4, 0.2, 2.7, and 1.5 mm, and the corresponding 3DPW errors by 1.5, 1.7, 0.6, and 3.3 mm. On the cross-age CMU sequences, Root MPJPE decreases from 100.2 to 79.2 mm (21.0%), while Joint-PA MPJPE decreases from 156.3 to 79.6 mm (49.1%). Figure 6 in the appendix shows CAHI4D directly alongside the compared methods for Hi4D close interactions and CMU toddler reconstruction. Joint-PA MPJPE retains inter-person placement under a shared alignment, so its larger reduction is consistent with improved relative geometry on the evaluated adult–child sequences.

Qualitative analysis. Figure 4 compares CAHI4D with existing methods on contact-rich adult–adult and adult–child interactions. Under severe mutual occlusion, large morphology differences, and close physical contact, CAHI4D produces more coherent body scale and relative placement across both input and novel views. In the shown examples, it preserves the intended contact with fewer visible interpenetrations. Appendix A.8 shows further frames from the Hi4D, CMU, and in-the-wild examples (Figure 7) and a longer 3DPW sequence (Figure 8). Figure 5 compares CAHI4D and JOSH on an in-the-wild adult–child video across multiple frames from oblique and side viewpoints.

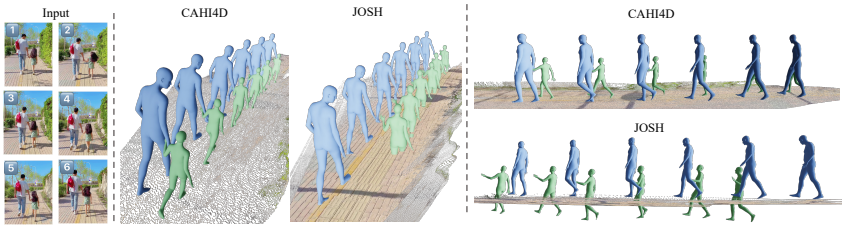

Figure 5: **Dynamic comparison of CAHI4D and JOSH.** Six frames of an in-the-wild adult–child sequence, with inputs and reconstructions from oblique and side viewpoints.

These observations complement the quantitative improvements in individual reconstruction and inter-person structure.

4.3 Ablation Studies

Cumulative framework ablation. The cumulative transition from VIMO initialization to stage (2) (Table 1) reduces CMU Root/Joint-PA MPJPE by 59.7/193.7 mm and Hi4D/3DPW Inter by 33.6/124.3 mm. DIGA uses depth to fit the scene-supported adult, validated interaction to place the

child, and the resulting projection-size change to correct scale. Relative to the shared attributes-only/no-DIGA baseline, stage (4) reduces CMU errors by 87.9/83.5 mm (Table 2). ACSF fits age-related proportions at fixed placement and scale, lowering CMU errors to 79.3/79.8 mm, although Hi4D Inter rises from 73.3 to 74.2 mm. IMRF refines motion under contact and temporal constraints with the geometric anchor fixed, lowering Inter by 3.1 mm to 71.1 mm on close-interaction Hi4D and by 14.0 mm to 93.1 mm on 3DPW.

Geometric anchoring ablation. Table 2 reports Root/Joint-PA MPJPE on CMU Panoptic. In the frame shown in Appendix Figure 9, the child has visible floor support and a reliable depth cue. In the w/o Interaction variant, depth and scene evidence still constrain placement (82.4/82.3 mm); in w/o Depth, validated interaction and scene evidence remain (79.4/84.2 mm). Both variants look close to Ours (79.2/79.6 mm), although Ours better reproduces the adult–child hand contact. By contrast, w/o DIGA visibly misestimates the child's size and placement, yielding 167.1/163.1 mm. This row shares the attributes-only configuration in Table 3 and removes both anchoring and downstream geometric shape evidence.

Table 2: Geometric anchoring ablation on the CMU Panoptic adult–child subset.

Variant	Root MPJPE↓	Joint-PA MPJPE↓
w/o DIGA	167.1	163.1
w/o Depth	79.4	84.2
w/o Interaction	82.4	82.3
Ours (w/ DIGA)	**79.2**	**79.6**

Shape-fitting strategy ablation. Table 3 compares shape cues on CMU Panoptic. Attributes only has no geometric cues and corresponds to w/o DIGA in Table 2, yielding 167.1/163.1 mm Root/Joint-PA MPJPE. Geometry only, which uses DIGA, lowers these errors to 79.4/79.9 mm, a reduction of 87.7/83.2 mm. Combining both gives 79.2/79.6 mm, a further numerical reduction of 0.2/0.3 mm. Thus, geometry accounts for most of the joint-error reduction; however, the Geometry only example in Appendix Figure 9 shows that a child can still be fitted with adult-like proportions at a shorter stature without attributes.

Table 3: Shape-fitting strategy ablation on the CMU Panoptic adult–child subset.

Variant	Root MPJPE↓	Joint-PA MPJPE↓
Attributes only	167.1	163.1
Geometry only	79.4	79.9
Ours (Both)	**79.2**	**79.6**

Interaction correction ablation. Table 4 compares material-contact loss and KTIC on Hi4D. All variants share anchored geometry, event-admission rules, and the remaining motion terms. Adding material-contact loss changes MPJPE/Inter/PD-MAE from 60.0/72.7/0.878 to 59.6/72.1/0.834. Independently solved chain targets give 59.7/72.6/0.834, with no further improvement. Replacing them with KTIC's shared-DoF coordination reduces the errors to 55.1/71.1/0.808, lower by 4.6/1.5/0.026 than independent targets.

Table 4: Interaction correction ablation on Hi4D.

Variant	MPJPE↓	Inter↓	PD-MAE↓
w/o $\mathcal{L}_{\mathrm{mat}}$ and KTIC	60.0	72.7	0.878
w/o KTIC	59.6	72.1	0.834
Independent chain targets	59.7	72.6	0.834
Ours (w/ KTIC)	**55.1**	**71.1**	**0.808**

5 Conclusion

We presented CAHI4D for monocular cross-age human interaction 4D reconstruction in a shared 3D space. DIGA combines validated interaction cues with depth to refine relative placement and derives sequence-consistent scale corrections from the resulting projection change. Shape is then fitted under the fixed geometric anchor before interaction-motion residuals are optimized. Within that motion stage, KTIC maps directed surface corrections to participant-specific kinematic chains and coordinates corrections affecting shared joint degrees of freedom, while contact and temporal terms constrain the same update. Without an admitted interaction relation, CAHI4D falls back to each participant's own evidence; without metric calibration, geometry remains in a relative shared gauge. Across the evaluated protocols, CAHI4D records the lowest errors among the compared methods, with its largest relative gains on CMU toddler.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing portions of the manuscript, identifying potentially relevant prior work, and providing feedback on research methodology and experimental design. CAHI4D also uses Qwen3-VL within the reconstruction pipeline to propose interaction hypotheses and estimate body attributes; interaction hypotheses are geometrically and temporally validated before use, as described in Section 3.1. The authors reviewed AI-assisted content and suggestions and take responsibility for the final text, claims, citations, results, and artifacts.

REFERENCES

Paolo Baerlocher and Ronan Boulic. An inverse kinematics architecture enforcing an arbitrary number of strict priority levels. *The Visual Computer*, 20(6):402–417, 2004. doi: 10.1007/s00371-004-0244-4.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

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 *Computer Vision – ECCV 2024*, pp. 202–218. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73337-6_12.

Laura Bravo-Sánchez, Matthieu Armando, Romain Brégier, Grégory Rogez, Serena Yeung-Levy, and Fabien Baradel. Anny-fit: All-age human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 3677–3687, 2026.

Romain Brégier, Fabien Baradel, Thomas Lucas, Salma Galaaoui, Matthieu Armando, Philippe Weinzaepfel, and Grégory Rogez. CondiMen: Conditional multi-person mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3880–3890, 2025.

Romain Brégier, Guénolé Fiche, Laura Bravo-Sánchez, Thomas Lucas, Matthieu Armando, Philippe Weinzaepfel, Grégory Rogez, and Fabien Baradel. Human mesh modeling for anny body. In *Computer Vision – ECCV 2026*, pp. 170–188. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37314-4_10.

Georgios Chatzichristodoulou, Niki Efthymiou, Panagiotis Filntisis, Georgios Pavlakos, and Petros Maragos. Age-inclusive 3D human mesh recovery for action-preserving data anonymization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5553–5563, 2026.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *International Conference on Learning Representations*, 2026. arXiv:2510.06219.

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3d interaction reasoning from 2d foundational models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22605–22615, 2025. doi: 10.1109/CVPR52734.2025.02105.

Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-HMR 2: Multi-person camera-centric human detection, mesh recovery and tracking. In *Computer Vision – ECCV 2026*, pp. 281–299. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37013-6_16. arXiv:2606.14841.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7214–7223, 2020. doi: 10.1109/CVPR42600.2020.00724.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):10870–10881, 2025. doi: 10.1109/TPAMI.2025.3601974.

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.

Nikolas Hesse, Sergi Pujades, Michael J. Black, Michael Arens, Ulrich G. Hofmann, and A. Sebastian Schroeder. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2540–2551, 2020. doi: 10.1109/TPAMI.2019.2917908.

Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1011–1021, 2024. doi: 10.1109/CVPR52733.2024.00102.

Buzhen Huang, Chen Li, Chongyang Xu, Dongyue Lu, Jinnan Chen, Yangang Wang, and Gim Hee Lee. Reconstructing close human interaction with appearance and proxemics reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17475–17485, 2025. doi: 10.1109/CVPR52734.2025.01628.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Hongbo Kang, Tianyi Zhou, Qingyang Yang, Hongwei Wen, Jing Huang, Yu-Kun Lai, and Kun Li. Crowd4D: Scene-aware monocular 4D crowd reconstruction. In *International Conference on Machine Learning*, 2026.

Gwanghyun Kim, Junghun James Kim, Suh Yoon Jeon, Jason Park, and Se Young Chun. Human interaction-aware 3D reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21252–21261, 2026.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. In *International Conference on Learning Representations*, 2026. arXiv:2501.02158.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013.

Lea Müller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3D social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9687–9697, 2024. doi: 10.1109/CVPR52733.2024.00925.

Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13468–13478, 2021. doi: 10.1109/CVPR46437.2021.01326.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019. doi: 10.1109/CVPR.2019.01123.

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, 2024. URL `https://arxiv.org/abs/2408.00714`.

Jun Saito, Jiefeng Li, Michael de Ruyter, Miguel Guerrero, Edy Lim, Ehsan Hassani, Roger Blanco Ribera, Hyejin Moon, Magdalena Dadela, Marco Di Lucca, Qiao Wang, Xueting Li, Jan Kautz, Simon Yuen, and Umar Iqbal. SOMA: Unifying parametric human body models, 2026. arXiv:2603.16858.

Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16094–16104, 2021. doi: 10.1109/CVPR46437.2021.01583.

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.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024.

Chentao Song, He Zhang, Haolei Yuan, Haozhe Lin, Jianhua Tao, Hongwen Zhang, and Tao Yu. MetricHMSR: Metric human mesh and scene recovery from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Sanjay Subramanian, Evonne Ng, Lea Müller, Dan Klein, Shiry Ginosar, and Trevor Darrell. Pose priors from language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7125–7135, 2025. doi: 10.1109/CVPR52734.2025.00668.

Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13243–13252, 2022. doi: 10.1109/CVPR52688.2022.01289.

Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. In *International Conference on 3D Vision*, pp. 53–63, 2021. doi: 10.1109/3DV53792.2021.00016.

Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. MultiPhys: Multi-person physics-aware 3D motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2331–2340, 2024.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Learning single-image calibration with geometric optimization. In *Computer Vision – ECCV 2024*, pp. 1–20, 2024.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision*, pp. 601–617, 2018. doi: 10.1007/978-3-030-01249-6_37.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kaiwen Wang, Kaili Zheng, Yiming Shi, Chenyi Guo, and Ji Wu. Towards metric-aware multi-person mesh recovery by jointly optimizing human crowd in camera space, 2025a.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *Computer Vision – ECCV 2024*, pp. 467–487. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73247-8_27. arXiv:2403.17346.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1148–1159, 2025b. doi: 10.1109/CVPR52734.2025.00115.

Hao Wen, Hongbo Kang, Jian Ma, Jing Huang, Yuanwang Yang, Haozhe Lin, Yu-Kun Lai, and Kun Li. DyCrowd: Towards dynamic crowd reconstruction from a large-scene video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3600465.

Zhenzhen Weng, Laura Bravo-Sánchez, Zeyu Wang, Christopher Howard, Maria Xenochristou, Nicole Meister, Angjoo Kanazawa, Arnold Milstein, Elika Bergelson, Kathryn L. Humphreys, Lee M. Sanders, and Serena Yeung-Levy. Artificial intelligence–powered 3d analysis of video-based caregiver–child interactions. *Science Advances*, 11(8):eadp4422, 2025. doi: 10.1126/sciadv.adp4422.

Qi Xia, Peishan Cong, Ziyi Wang, Yujing Sun, Qin Sun, Xinge Zhu, Mao Ye, Ruigang Yang, and Yuexin Ma. SocialMirror: Reconstructing 3D human interaction behaviors from monocular videos with semantic and geometric guidance, 2026. arXiv:2604.13581.

Chongyang Xu, Buzhen Huang, Chengfang Zhang, Ziliang Feng, and Yangang Wang. Adapting human mesh recovery with vision-language feedback, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35: 38571–38584, 2022.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21222–21232, 2023. doi: 10.1109/CVPR52729.2023.02033.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17016–17027, 2023. doi: 10.1109/CVPR52729.2023.01632.

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.

Yufei Zhang, Hanjing Wang, Jeffrey O. Kephart, and Qiang Ji. Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9020–9032, 2023. doi: 10.1109/ICCV51070.2023.00828.

Kaili Zheng, Kaiwen Wang, Xun Zhu, Chenyi Guo, and Ji Wu. InterMesh: Explicit interaction-aware end-to-end multi-person human mesh recovery, 2026. arXiv:2605.04554.

APPENDIX

The appendix provides the VLM prompt and additional method details, followed by the CMU evaluation protocol, quantitative and qualitative results, and limitations.

Throughout the appendix, $\varepsilon = 10^{-6}$ is the numerical floor used in denominators and logarithmic ratios.

A.1 VLM KEYFRAME INFERENCE AND PROMPT

We select candidate keyframes where the tracked participants' 2D poses are close enough to suggest an interaction. Qwen3-VL receives these frames in temporal order with track IDs, masks, joint overlays, frame IDs, and fixed SOMA vocabularies. The surface regions are `pelvis`, `torso`, `head`, and bilateral `upper_arm`, `forearm`, `hand`, `thigh`, `shin`, and `foot`.

> *Prompt template.* Analyze `[ORDERED_KEYFRAMES]` in order. Track identities using `[TRACK_IDS]` and `[IDENTITY_LEGEND]`; use only `[FRAME_IDS]`, `[JOINTS]`, and the supplied SOMA surface regions. Return valid JSON with `global_description`, `attributes`, `joint_states`, `pair_hypotheses`, and `directed_support`; include confidence in $[0, 1]$ and supporting frame IDs. Return no explanatory text.
>
> For each person, report an `age_interval`, `body_build` (`slim`, `average`, `heavy`, or `uncertain`), and evidence-supported `head_stature_ratio`, `shoulder_hip_ratio`, `arm_torso_ratio`, and `leg_torso_ratio` intervals. If a suitable reference person is visible, report a broad `height_interval` as target-to-reference stature ratio and its `reference_person_id`. Set unsupported age or height intervals to null and omit unsupported body ratios.
>
> For each supplied frame, person, and joint, label `joint_states` as `no_evidence`, `visible_candidate`, `occluded_candidate`, or `uncertain`. For each plausible cross-person contact, add a `pair_hypotheses` entry with a person ID, SOMA joint, and surface region at each endpoint; label its `relation` as `contact`, `support`, `holding_hands`, `hugging`, or `leaning_on`, and `visibility` as `direct`, `occluded`, or `uncertain`. For plausible load-bearing pairs, report `supporter_id`, `supported_id`, IDs of pairs labeled `support`, and `ground_state` (`grounded`, `partially_off_ground`, `off_ground`, or `uncertain`).
>
> Use neighboring keyframes, limb configuration, and the overall action to assess occluded contact; image overlap alone is insufficient. Lower confidence or abstain when evidence is weak. Do not infer metric stature, exact 3D contact locations, or unobserved contact timing.

The output structure is illustrated below. Placeholder values indicate types; IDs and frame references must come from the supplied inputs.

```
{
  "global_description": {"interaction": "",
    "person_person_interaction": true,
    "person_scene_support": false, "confidence": 0.0,
    "evidence_frame_ids": []},
  "attributes": [{
    "person_id": "", "age_interval": null,
    "body_build": {"category": "uncertain",
      "confidence": 0.0, "evidence_frame_ids": []},
    "height_interval": {"reference_person_id": "",
      "lower": 0.0, "upper": 0.0,
      "confidence": 0.0, "evidence_frame_ids": []},
    "ratio_intervals": [{"attribute": "head_stature_ratio",
      "lower": 0.0, "upper": 0.0,
      "confidence": 0.0, "evidence_frame_ids": []}]
  }],
  "joint_states": [{
    "frame_id": 0, "person_id": "", "joint": "",
    "state": "no_evidence", "confidence": 0.0
  }],
  "pair_hypotheses": [{
    "pair_id": "", "person_a_id": "", "joint_a": "",
```

```
    "region_a": "", "person_b_id": "", "joint_b": "",
    "region_b": "", "relation": "contact",
    "visibility": "direct", "confidence": 0.0,
    "evidence_frame_ids": [], "evidence": ""
  }],
  "directed_support": [{
    "supporter_id": "", "supported_id": "",
    "supporting_pair_ids": [], "ground_state": "uncertain",
    "confidence": 0.0, "evidence_frame_ids": [],
    "evidence": ""
  }]
}
```

Age, build, and body-ratio estimates constrain only the corresponding participant's feasible Anny phenotype. Per-joint states record visible or occluded evidence for proposed relations; only validated relations constrain optimization. Relative height is a weak appearance cue; metric scale requires an independent validated source. We reject invalid IDs, joint or region names, frame references, and support-pair references. Geometric and temporal checks determine which contact and support hypotheses are admitted. A VLM supporter–supported proposal does not determine DIGA's reference–target direction $i \to j$ or KTIC's correction-owner direction $o_e \leftarrow q_e$.

A.2 DIGA: From Placement to Geometric Anchoring

A.2.1 Semantic cues and reference graph.

The relative geometric reliability of participant p is

$$a_p^{\mathrm{rel}} = \frac{\sum_t v_{p,t}(c_{p,t}^{\mathrm{mask}} + c_{p,t}^{\mathrm{depth}} + c_{p,t}^{\mathrm{scene}})}{3\sum_t v_{p,t} + \varepsilon}, \tag{12}$$

where $v_{p,t} \in \{0,1\}$ marks a valid frame and the three confidence terms in $[0,1]$ measure calibrated mask, depth, and scene reliability. Visible-joint evidence enters the image residual of the subsequent reference check, rather than this reliability score.

The Anny adapter maps phenotype $\boldsymbol{\phi}_p \in \mathcal{D}_{\mathrm{A}}$ to age and scale-invariant body ratios $F_m(\boldsymbol{\phi}_p)$, where m indexes head-to-stature, shoulder-to-hip, arm-to-torso, and leg-to-torso ratios. The attribute-supported set $\Phi_p \subseteq \mathcal{D}_{\mathrm{A}}$ intersects age and ratio intervals with confidence at least 0.6 in descending confidence order; an interval is discarded if it makes the intersection empty, and $\Phi_p = \mathcal{D}_{\mathrm{A}}$ if none is retained. Phenotype distances use normalized control coordinates. These intervals guide morphology, not metric scale. VLM-proposed support relations first define candidate components. In each component, a participant with $a_p^{\mathrm{rel}} \geq 0.6$ is selected as a candidate reference. After scene-scale and monocular-depth alignment to this reference and fitting it from its own evidence, the reference must pass its residual check; only relations passing surface-distance, normal, image-overlap, aligned-depth-order, and temporal-persistence checks enter the directed DIGA reference graph. Unsupported or cyclic edges are excluded. The check scores each candidate by the median over valid frames of its mean clipped, normalized pre-Huber image, depth, and scene residuals; when at least three candidates are available, it applies a cross-candidate gate using 2.5 times the median absolute deviation (MAD) of candidate scores. Metric calibration requires a separate validated length source. A component without a reliable root uses its own evidence. For placement, we fix

$$\boldsymbol{\phi}_j^0 = \operatorname*{argmin}_{\boldsymbol{\phi} \in \Phi_j} \|\boldsymbol{\phi} - \boldsymbol{\phi}_j^{\mathrm{init}}\|_2^2. \tag{13}$$

This phenotype remains fixed through placement and geometric anchoring; shape fitting updates it later.

A.2.2 Placement and support anchors.

Let Π_t be the fixed camera projection and $\widehat{M}_{p,t}^{\mathrm{vis}}$ the rendered visible-person mask. On the original uncropped canvas, both observed and rendered masks use $\mathrm{Extent}(M) = \sqrt{\sum_{\mathbf{u}} M(\mathbf{u})}$ over valid pixels; their extents are $h_{p,t}^{\mathrm{img}}$ and $\widehat{h}_{p,t}$. We use the Huber penalty $\rho(x) = \frac{1}{2}x^2$ for $|x| \leq 0.05$ and $0.05(|x| - 0.025)$ otherwise. The extent term is

$$\widehat{h}_{p,t} = \mathrm{Extent}(\widehat{M}_{p,t}^{\mathrm{vis}}), \qquad \mathcal{L}_{\mathrm{ext}} = \sum_{p,t} c_{p,t}^{\mathrm{ext}} \rho\left(\log \frac{\widehat{h}_{p,t} + \varepsilon}{h_{p,t}^{\mathrm{img}} + \varepsilon}\right), \tag{14}$$

where $c^{\text{ext}}_{p,t} \in [0,1]$ is the calibrated extent confidence. After selecting a scene-supported reference, we use its validated scene correspondence to align scene scale in the current shared gauge. Monocular inverse depth is then aligned to the reference depth in that gauge and frozen for subsequent depth residuals; without an independent metric calibration, only ordinal and pairwise differences constrain relative depth.

VIMO root orientation and position are mapped through the frozen camera extrinsics before DIGA. Let $\ell^0_{p,t}$ be the detached initialized body-box diagonal. After fitting the reference root and orientation from its own evidence, we freeze it and find closest surface correspondences within each candidate support body-part pair. These yield reference points $\mathbf{c}_{i_k,k,t}$, normals $\mathbf{n}_{i_k,k,t}$, and target material points $\mathbf{z}_{j,k,t}$. With $\bar{\ell}^0_{p,k} = \text{median}_{t\in\mathcal{I}_k} \ell^0_{p,t}$, the frozen reference anchor is

$$\delta_k = \text{clip}_{[0,\,0.02\min(\bar{\ell}^0_{i_k,k},\bar{\ell}^0_{j,k})]} \underset{t\in\mathcal{I}_k}{\text{median}} \left[\mathbf{n}^\top_{i_k,k,t}(\mathbf{z}_{j,k,t} - \mathbf{c}_{i_k,k,t})\right], \qquad \mathbf{a}_{i_k,k,t} = \mathbf{c}_{i_k,k,t} + \delta_k \mathbf{n}_{i_k,k,t}.$$

Reference anchors and target triangle/barycentric identities remain fixed through placement, anchoring, and shape fitting. Evaluating a target material identity on the pelvis-centered mesh at phenotype $\boldsymbol{\phi}_j$ defines $\overline{\mathbf{q}}_{j,k,t}(\boldsymbol{\phi}_j)$ in Eq. 3. With calibrated support confidence $c^{\text{sup}}_{j,t} \in [0,1]$, the placement residual is

$$\mathcal{L}^{\text{anc}}_{\text{support}}(\mathbf{r}_j, s) = \sum_{t:\mathcal{K}_{j,t}\neq\varnothing} c^{\text{sup}}_{j,t}\, \rho\left(\frac{\|\mathbf{r}_{j,t} - \widehat{\mathbf{r}}^{\text{sup}}_{j,t}(\boldsymbol{\phi}^0_j, s)\|_2}{\ell^0_{j,t} + \varepsilon}\right). \tag{15}$$

Reference fitting uses the image, depth, extent, scene, and DIGA temporal residuals in Table 5 with $\mathbf{r}^{\text{geo}}$ and $\mathbf{R}$ active. Target placement adds Eq. 15 at $s^0_j = 1$ and updates only $\mathbf{r}^{\text{geo}}_{j,1:T}$.

A.2.3 Geometric anchor refinement.

Valid ratio frames $\mathcal{T}_j$ require untruncated isolated amodal renders, positive projected sizes in Eq. 4, admitted support, and agreement with the target's own visible-mask or torso-joint evidence. The weights $w_{j,t}$ combine image, support, and projection confidence. With fewer than $\max(5, \lceil 0.1T \rceil)$ valid frames, we retain $s^0_j = 1$. The robust estimate $\widehat{\eta}_j$ in Eq. 6 initializes the bounded full-perspective refinement:

$$\eta^\star_j = \underset{\eta\in[\log 0.5, \log 2]}{\text{argmin}} \sum_{t\in\mathcal{T}_j} w_{j,t}\, \rho\left(\log \frac{\mathcal{H}_{j,t}(\mathbf{r}^{\text{geo}}_{j,t}(s^0_j e^\eta), s^0_j e^\eta) + \varepsilon}{h^-_{j,t} + \varepsilon}\right) + \lambda_\eta(\eta - \widehat{\eta}_j)^2. \tag{16}$$

At each trial ratio $s = s^0_j e^\eta$, Eq. 7 updates supported-frame placement and a smooth extension supplies other frames. We evaluate the temporarily scaled initialized mesh of Eq. 8, never a candidate final shape. A trial is accepted only if projection agreement improves, support and image/scene residuals remain valid, and $sH_{\text{A}}(\boldsymbol{\phi}^0_j)$ is attainable within $\mathcal{D}_{\text{A}}$. Otherwise the step is shortened; if none passes, $s^\star_j = s^0_j$ and $\mathbf{r}^{\text{geo}\star}_{j,t} = \mathbf{r}^+_{j,t}$. The visible-mask extent in Eq. 14 checks image consistency but does not replace the isolated amodal statistic in Eq. 4.

A.3 Shape-Then-Interaction Optimization

A.3.1 Anchor-conditioned shape refinement.

For an underconstrained target, shape fitting fixes $\mathbf{r}^{\text{geo}\star}$, $\mathbf{R}$, $\boldsymbol{\theta}^0$, camera, scene, and reference anchors. The temporary ratio supplies only the target stature

$$h^{\text{anc}}_j = s^\star_j H_{\text{A}}(\boldsymbol{\phi}^0_j), \tag{17}$$

Shape evidence is evaluated on the unscaled candidate mesh

$$\mathbf{X}^{\text{shape}}_{j,t}(\boldsymbol{\phi}_j) = \mathbf{r}^{\text{geo}\star}_{j,t} + \mathbf{R}_{j,t}\overline{\mathbf{Y}}_{j,t}(\boldsymbol{\phi}_j, \boldsymbol{\theta}^0_{j,t}). \tag{18}$$

With $\ell^{\text{anc}}_{j,t}$ the detached anchor body-box diagonal, support consistency uses the candidate's material point at ratio one:

$$\mathcal{L}^{\text{shape}}_{\text{support}}(\boldsymbol{\phi}_j) = \sum_{t:\mathcal{K}_{j,t}\neq\varnothing} c^{\text{sup}}_{j,t}\, \rho\left(\frac{\|\mathbf{r}^{\text{geo}\star}_{j,t} - \widehat{\mathbf{r}}^{\text{sup}}_{j,t}(\boldsymbol{\phi}_j, 1)\|_2}{\ell^{\text{anc}}_{j,t} + \varepsilon}\right). \tag{19}$$

We write $\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{sil}}$ for the target's own image terms; aligned depth and projected extent also constrain its shape. The complete fit is

$$\begin{aligned}\boldsymbol{\phi}_j^\star = \underset{\substack{\boldsymbol{\phi}_j \in \mathcal{D}_{\text{A}} \\ H_{\text{A}}(\boldsymbol{\phi}_j) = h_j^{\text{anc}}}}{\operatorname{argmin}} \quad & \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{ext}}\mathcal{L}_{\text{ext}} \\ & + \lambda_{\text{sup}}\mathcal{L}_{\text{support}}^{\text{shape}} + \lambda_\phi \operatorname{dist}(\boldsymbol{\phi}_j, \Phi_j)^2 + \lambda_{\text{init}}\|\boldsymbol{\phi}_j - \boldsymbol{\phi}_j^0\|_2^2.\end{aligned} \tag{20}$$

The equality constraint preserves anchor-derived stature while allowing age-dependent proportions to change. We project phenotype updates onto the feasible stature level set in $\mathcal{D}_{\text{A}}$; infeasible proposals are rejected. Participants without shape refinement retain $\boldsymbol{\phi}_p^\star = \boldsymbol{\phi}_p^0$. The ratio $s_j^\star$ is not applied to the final mesh.

A.4 Kinematic-Topology-Guided Interaction Correction

A.4.1 Deterministic event construction.

Penetration candidates arise from visible overlap; material-contact candidates arise from overlap or admitted interaction hypotheses, including occluded support. In either case, surface distance, normal compatibility, and temporal persistence determine admission and correspondence rather than the VLM. Candidate overlap uses a three-pixel mask dilation; opposed surface normals must have dot product at most -0.3, and material-contact candidates require a score of at least 0.6. Let $\ell_{p,t}^\star$ be the detached fitted-base body-box diagonal and $\ell_{e,t}^\star = \min(\ell_{o_e,t}^\star, \ell_{q_e,t}^\star) > 0$. Penetration and material-contact events remain distinct sets, $\mathcal{E} = \mathcal{E}_{\text{pen}} \dot{\cup} \mathcal{E}_{\text{mat}}$.

Each penetration event uses a connected visible owner patch; each material event fixes an owner–target triangle pair and its barycentric coordinates at a representative valid frame t_e^0. Patch weights $\omega_{e,v}^{\text{patch}}$ sum to one. Valid frames form an admitted interval $\mathcal{I}_e$ of at least $\max(5, \lceil 0.1T \rceil)$ frames; material identities remain fixed throughout it, whereas penetration nearest points may be relinearized.

Skinning weights on $\mathcal{P}_e$ select the owner participant's affected joints and their non-root kinematic ancestors, giving $\mathcal{C}_e \subset \{(o_e, a) : a \text{ is a joint}\}$. We select the correction owner from directed-role evidence when available, pose-update eligibility, and relative geometric reliability. For overlap-only penetration, the role term is neutral; an ambiguous owner is rejected. The other participant is a detached target q_e, not a second gradient path. Event confidence $g_{e,t} \in [0, 1]$ combines geometric, temporal, and ownership evidence, is zero outside $\mathcal{I}_e$, and is detached within each motion-update block. Material contact uses the fixed target triangle's right-handed tangent frame $\mathbf{B}_{e,t}^q$; invalid material correspondences are discarded rather than rematched.

A.4.2 Surface requests and joint targets.

Let $\mathbf{x}_{o_e,t,v}$ be owner vertex v in shared space and let the fixed patch weights $\omega_{e,v}^{\text{patch}}$ sum to one. Writing $\boldsymbol{\xi}_{\mathcal{C}_e}$ for the chain's right-multiplicative local rotation coordinates, the owner point and participant-indexed surface Jacobian, evaluated at the current iterate $\boldsymbol{\xi}_{\mathcal{C}_e} = \mathbf{0}$, are

$$\mathbf{x}_{e,t} = \sum_{v \in \mathcal{P}_e} \omega_{e,v}^{\text{patch}} \mathbf{x}_{o_e,t,v}, \qquad \mathbf{J}_{e,t} = \frac{\partial \mathbf{x}_{e,t}}{\partial \boldsymbol{\xi}_{\mathcal{C}_e}}. \tag{21}$$

For either event type, the signed-gap target is $\delta_{e,t}^{\text{gap}} = 0.005\ell_{e,t}^\star$; material-contact admission at $0.02\ell_{e,t}^\star$ does not replace this target tolerance. A penetration event uses a detached nearest target point $\mathbf{y}_{e,t}^{\text{near}}$ and unit outward normal $\mathbf{n}_{e,t}^{\text{near}}$. A material event instead uses the fixed material coordinate and legal nonpenetrating target $\mathbf{z}_{e,t}^{\text{ref}}$ from Eq. 24. The type-specific clipped request and anisotropic weight are

$$\begin{aligned} d_{e,t}^{\text{pen}} &= [\delta_{e,t}^{\text{gap}} - (\mathbf{n}_{e,t}^{\text{near}})^\top(\mathbf{x}_{e,t} - \mathbf{y}_{e,t}^{\text{near}})]_+, \qquad e \in \mathcal{E}_{\text{pen}}, \\ \mathbf{b}_{e,t} &= \begin{cases} \operatorname{clip}_{\delta_{\max}}(d_{e,t}^{\text{pen}} \mathbf{n}_{e,t}^{\text{near}}), & e \in \mathcal{E}_{\text{pen}}, \\ \operatorname{clip}_{\delta_{\max}}\left(\mathbf{B}_{e,t}^q[\mathbf{z}_{e,t}^{\text{ref}} - \mathbf{z}_{e,t}^{\text{mat}}]\right), & e \in \mathcal{E}_{\text{mat}}, \end{cases} \\ \mathbf{W}_{e,t} &= w_n \mathbf{n}_{e,t}^{\text{near}}(\mathbf{n}_{e,t}^{\text{near}})^\top + w_\tau\left(\mathbf{I} - \mathbf{n}_{e,t}^{\text{near}}(\mathbf{n}_{e,t}^{\text{near}})^\top\right). \end{aligned} \tag{22}$$

Here $\mathrm{clip}_c(\mathbf{v}) = \min(1, c/\|\mathbf{v}\|_2)\mathbf{v}$ for $\mathbf{v} \neq \mathbf{0}$ and is zero otherwise, with $\delta_{\max} = 0.02\ell^\star_{e,t}$. For a material event, $\mathbf{n}^{\mathrm{near}}_{e,t}$ is its fixed target-triangle normal. The matrix $\mathbf{W}_{e,t}$ emphasizes the target normal with $(w_n, w_\tau) = (1, 0.1)$, and $\|\mathbf{z}\|^2_{\mathbf{W}} = \mathbf{z}^\top \mathbf{W}\mathbf{z}$. The target point, normal, confidence, and correspondences are detached within a motion-update block.

At frame t, $\mathcal{E}_t = \{e : t \in \mathcal{I}_e,\ g_{e,t} > 0\}$. Events sharing a participant-indexed joint belong to one component $\mathcal{Q}_t$, where Eq. 10 gives one increment per physical joint. Let $\mathcal{Q}_t(p, a)$ contain joint (p, a) and $\mathbf{S}_{p,a}$ select its coordinates. A bounded, energy-decreasing backtracking step with $\theta_{\max} = \pi/2$ gives $\Delta\boldsymbol{\xi}_{p,a,t} = \mathrm{clip}_{\theta_{\max}}(\alpha^{(n)}_{\mathcal{Q}_t(p,a),t}\mathbf{S}_{p,a}\boldsymbol{\xi}^\star_{\mathcal{Q}_t(p,a),t})$ and

$$\begin{aligned} \widehat{\mathbf{Q}}_{p,a,t} &= \mathbf{Q}^{(n)}_{p,a,t}\,\mathrm{Exp}([\Delta\boldsymbol{\xi}_{p,a,t}]_\times)\,, \\ \bar{g}_{p,a,t} &= 1 - \prod_{e\in\mathcal{E}_t:\,(p,a)\in\mathcal{C}_e} (1 - g_{e,t}). \end{aligned} \tag{23}$$

These targets are detached before entering Eq. 11.

A.5 Joint Interaction-Motion Objective and Optimization

A.5.1 Persistent contact and root trust.

For persistent contact, the admitted owner and target triangles define deformed material points $\mathbf{x}^{\mathrm{mat}}_{e,t}$ and $\mathbf{y}^{\mathrm{mat}}_{e,t}$. In the deterministic, stop-gradient target tangent frame $\mathbf{B}^q_{e,t}$, their relative material coordinate is

$$\begin{aligned} \mathbf{z}^{\mathrm{mat}}_{e,t} &= (\mathbf{B}^q_{e,t})^\top\big(\mathbf{x}^{\mathrm{mat}}_{e,t} - \mathrm{sg}(\mathbf{y}^{\mathrm{mat}}_{e,t})\big)\,, \\ \mathbf{z}^\tau_e &= \mathbf{P}_\tau\,\mathrm{sg}(\mathbf{z}^{\mathrm{mat},0}_{e,t^0_e}), \qquad \mathbf{z}^{\mathrm{ref}}_{e,t} = \mathbf{z}^\tau_e + \delta^{\mathrm{gap}}_{e,t}\mathbf{e}_3, \end{aligned} \tag{24}$$

Here $\mathbf{P}_\tau = \mathrm{diag}(1, 1, 0)$, $\mathbf{e}_3 = (0, 0, 1)^\top$, and $\mathbf{z}^{\mathrm{mat},0}_{e,t^0_e}$ evaluates the first line on the frozen fitted base state. Tangential coordinates preserve material identity; the normal coordinate enforces a non-penetrating gap. We use $\mathbf{D}_\tau = \mathrm{diag}(\beta_\tau, \beta_\tau, 1)$ with $\beta_\tau = 0.1$ in Eq. 27. For root updates, let $\mathcal{E}^{\mathrm{root}}_{p,t} = \{e \in \mathcal{E}_t : o_e = p,\ e \text{ permits root update}\}$. The horizontal trust region is

$$\begin{aligned} R^{\mathrm{root}}_{p,t} &= \begin{cases} \kappa_{\mathrm{root}}\min_{e\in\mathcal{E}^{\mathrm{root}}_{p,t}} d^{\mathrm{pair}}_{e,t}, & \mathcal{E}^{\mathrm{root}}_{p,t} \neq \varnothing \text{ and } p \text{ is not a reference}, \\ 0, & \text{otherwise}, \end{cases} \\ \|\mathbf{P}_h\Delta\mathbf{r}^{\mathrm{mot}}_{p,t}\|_2 &\le R^{\mathrm{root}}_{p,t}, \qquad \mathbf{v}_g^\top\Delta\mathbf{r}^{\mathrm{mot}}_{p,t} = 0. \end{aligned} \tag{25}$$

Here $\mathbf{v}_g$ is fixed unit gravity, $\mathbf{P}_h = \mathbf{I} - \mathbf{v}_g\mathbf{v}_g^\top$, $\kappa_{\mathrm{root}} = 0.05$, and $d^{\mathrm{pair}}_{e,t} = \|\mathbf{P}_h(\mathbf{r}^{\mathrm{geo}\star}_{o_e,t} - \mathbf{r}^{\mathrm{geo}\star}_{q_e,t})\|_2$ is the frozen horizontal owner–target distance. Projecting each eligible $\Delta\mathbf{r}^{\mathrm{mot}}_{p,t}$ onto this disk preserves Eq. 25; reference participants and frames without eligible events have zero root residual.

A.5.2 Residual definitions.

All residuals exclude missing observations and failed visibility tests. Let d_I be the image diagonal, sIoU soft intersection over union, D_t rendered inverse depth, u_t aligned observed inverse depth, and S_{scene} the frozen scene signed-distance field (positive outside). For uniform frame spacing, define

$$\boldsymbol{\omega}^R_{p,t} = \mathrm{Log}((\mathbf{R}_{p,t-1})^\top\mathbf{R}_{p,t})^\vee, \qquad \boldsymbol{\omega}^Q_{p,a,t} = \mathrm{Log}((\mathbf{Q}_{p,a,t-1})^\top\mathbf{Q}_{p,a,t})^\vee.$$

Both use the principal logarithm; proposals crossing its $\pi - 10^{-4}$ angle guard are rejected. Throughout, displayed residual sums omit division by their valid-term count (and equal zero when empty); weighted sums of loss terms are not re-normalized. Let c_{pvt} denote depth-sample confidence and $\sigma_{u,t}$ the robust inverse-depth uncertainty estimated from validated pixels. The terms are:

The motion-temporal term uses the resulting rotations $\mathbf{Q}$, not the raw residual vectors. For uncalibrated inverse depth, the ordinal hinge penalizes rendered sample pairs whose depth ordering disagrees with the frozen observation; no metric depth target is introduced.

Table 5: Residuals supplementing the objectives in Section 3. Δ^2 is a centered Euclidean second difference.

Term	Valid residual before averaging
$\mathcal{L}_{\text{2D}}$	$c_{pjt}\rho(\|\Pi_t(\mathbf{x}^{\text{joint}}_{p,j,t}) - \mathbf{u}_{pjt}\|_2/d_I)$ for every visible joint.
$\mathcal{L}_{\text{sil}}$	$1 - \text{sIoU}(\widehat{M}_{p,t}, M_{p,t})$ over unoccluded rendered pixels.
$\mathcal{L}_{\text{depth}}$	$c_{pvt}\rho((D_t(\Pi_t(\mathbf{X}_{p,t,v})) - u_t(\Pi_t(\mathbf{X}_{p,t,v})))/\sigma_{u,t})$ when calibrated; otherwise an ordinal hinge for each cached ordered pair.
$\mathcal{L}_{\text{scene}}$	$\rho([-S_{\text{scene}}(\mathbf{X}_{p,t,v})]_+/\ell^0_{p,t})$ for penetrating vertices plus $c^{\text{floor}}\rho(\|S_{\text{scene}}(\mathbf{x}^{\text{foot}})\|/\ell^0_{p,t})$ for validated support feet.
$\mathcal{L}_{\text{DIGA}-\text{temp}}$	$\|\Delta^2\mathbf{r}^{\text{geo}}_{p,t}\|_2^2/(\ell^0_{p,t})^2 + \|\boldsymbol{\omega}^R_{p,t+1} - \boldsymbol{\omega}^R_{p,t}\|_2^2$.
$\mathcal{L}^{\text{mot}}_{\text{image}}$	The $\mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{sil}}$ rows re-evaluated with only $\Delta\boldsymbol{\theta}^{\text{mot}}$ and $\Delta\mathbf{r}^{\text{mot}}$ active.
$\mathcal{L}_{\text{root}}$	$\|\mathbf{P}_h\Delta\mathbf{r}^{\text{mot}}_{p,t}\|_2^2/(\ell^\star_{p,t})^2$, subject to Eq. 25.
$\mathcal{L}_{\text{motion}-\text{temp}}$	$\sum_a \|\boldsymbol{\omega}^Q_{p,a,t+1} - \boldsymbol{\omega}^Q_{p,a,t}\|_2^2 + \|\Delta^2\Delta\mathbf{r}^{\text{mot}}_{p,t}\|_2^2/(\ell^\star_{p,t})^2$.

A.5.3 Joint interaction-motion objective.

Penetration events relinearize nearest points; persistent contact keeps the fixed triangle pair and barycentric coordinates. Their losses are

$$\mathcal{L}_{\text{pen}} = \sum_{e\in\mathcal{E}_{\text{pen}}}\sum_{t\in\mathcal{I}_e} g_{e,t}\rho\left(\frac{d^{\text{pen}}_{e,t}}{\ell^\star_{e,t}}\right), \tag{26}$$

$$\mathcal{L}_{\text{mat}} = \sum_{e\in\mathcal{E}_{\text{mat}}}\sum_{t\in\mathcal{I}_e} g_{e,t}\rho\left(\frac{\|\mathbf{z}^{\text{mat}}_{e,t} - \mathbf{z}^{\text{ref}}_{e,t}\|_{\mathbf{D}_\tau}}{\ell^\star_{e,t}}\right). \tag{27}$$

The interaction-motion stage jointly optimizes the residuals in Section 3.1 with KTIC targets and image, penetration, material-contact, root, and temporal terms:

$$\begin{aligned}\mathcal{L}_{\text{motion}} = \mathcal{L}^{\text{mot}}_{\text{image}} + \lambda_{\text{KTIC}}\mathcal{L}_{\text{KTIC}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{mat}}\mathcal{L}_{\text{mat}} \\ + \lambda_{\text{root}}\mathcal{L}_{\text{root}} + \lambda_{\text{motion}-\text{temp}}\mathcal{L}_{\text{motion}-\text{temp}}.\end{aligned} \tag{28}$$

A.5.4 Optimization settings.

CAHI4D uses PyTorch and optimizes each video independently. Adam uses learning rates of $0.0025/0.002/0.00015$ and $80/80/80$ steps for DIGA/shape/interaction optimization, respectively; scale optimization uses 0.001, and bounded root refinement uses 0.003. DLS uses a height-normalized base damping coefficient of 0.002. Main loss weights are $30/20/25/10$ (DIGA mask depth, horizontal mask alignment, projected extent, 3D anchors), $6/2/0.4$ (shape prior, mask, mask Chamfer), and $3/3/20/0.5$ (base reconstruction objective, KTIC, penetration, contact band), with graph regularization weighted by 0.05. In DIGA, the interpersonal contact loss has a weight of 0.25, with a coefficient of 0.35 for its contact gap penalty; contact-event anchors and relative stationarity constraints are weighted by 50 and 10, respectively. In the KTIC runtime ablation, DLS computation and its associated gradient projection are disabled; inputs and other optimization settings stay fixed.

A.6 CMU Selection Protocol

The 13 monocular adult–child clips (2,078 frames) come from five toddler sequences: `160906_ian1`, `160906_ian2`, `160906_ian3`, `160906_ian5`, and `170915_toddler5`. We select one available view per clip in which both participants have valid 3D joint annotations, excluding views with severe occlusion, out-of-frame participants, or missing annotations. CMU provides no SMPL parameters. All compared methods use the same selected views and frames. This selected-view protocol differs from Anny-Fit's published two-camera evaluation (Bravo-Sánchez et al., 2026).

A.7 Additional Quantitative Comparison

Figure 6 contrasts close-interaction performance on Hi4D with reconstruction accuracy on the selected CMU toddler views and frames used in Table 1. The CMU panel focuses on the four methods that support toddler reconstruction.

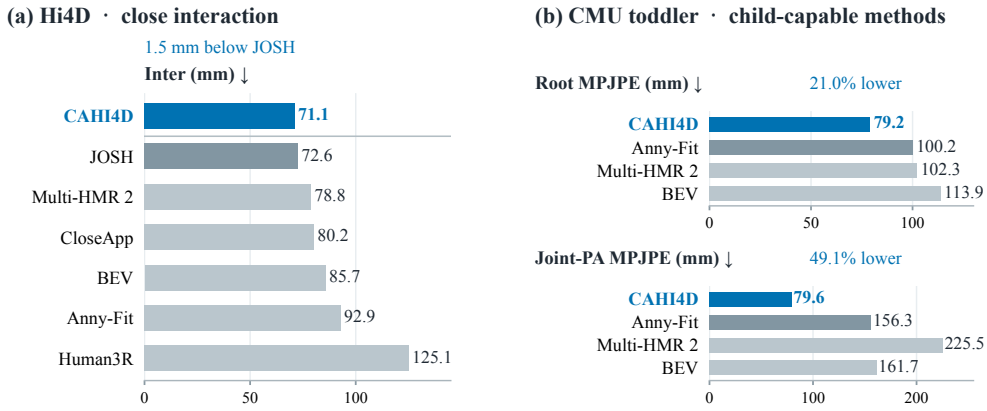

Figure 6: **Close and cross-age interaction.** (a) Hi4D inter-person joint-distance error (Inter) across CAHI4D and six baselines. (b) CMU Root and Joint-PA MPJPE for the four toddler-capable methods on matched views and frames; percentages compare CAHI4D with Anny-Fit. Bars show error in mm; lower is better. CAHI4D is highlighted in blue.

A.8 Qualitative Visualizations

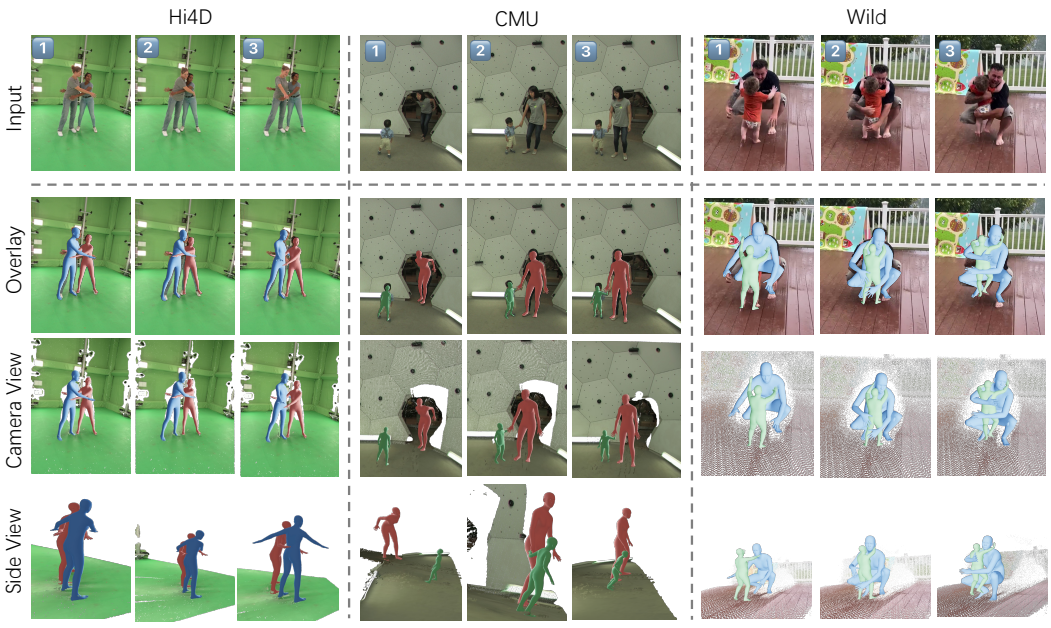

Figure 7: **Multi-frame qualitative results.** Three frames from each Hi4D, CMU, and in-the-wild sequence are shown as input, overlay, camera-view, and side-view rows.



Figure 8: **Long-sequence 3DPW reconstruction.** Five selected video frames (top) and CAHI4D reconstructions of both people in a shared scene view (bottom). Numbered markers link corresponding time points.

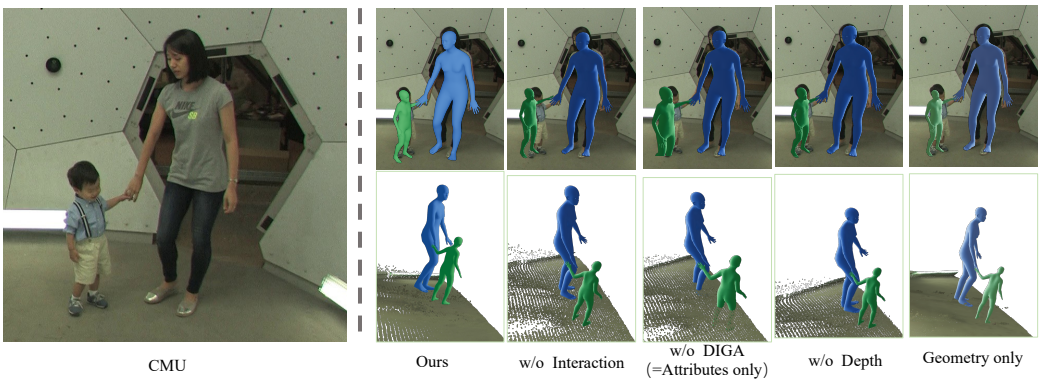

Figure 9: **Qualitative geometric and shape ablations on CMU Panoptic.** An input frame (left), input-view overlays (top), and a shared side view (bottom) compare the variants in Tables 2 and 3.

A.9 Runtime Analysis

Table 6: Fixed-step optimization runtime on a 100-frame Hi4D pose37 clip.

Stage	w/ KTIC	w/o KTIC
DIGA	52.46 s	52.46 s
Shape fitting	65.07 s	65.07 s
Interaction-motion fitting	311.31 s	180.01 s
Total optimization	428.83 s	297.54 s
Optimization FPS	0.233	0.336

We time one two-person clip on a single NVIDIA GeForce RTX 4090 D GPU. Both variants use 80 interaction-motion steps; all other settings and the DIGA/shape timings are shared. Stages are timed separately. FPS divides 100 frames by total optimization time; initialization, evaluation, and post-processing are excluded.

A.10 Limitations

CAHI4D relies on initialized bodies and predicted 2D keypoints, depth, and scale-invariant attributes. Occlusion, low resolution, and extreme viewpoints weaken these cues; erroneous predictions that pass validation can bias placement or shape. Weak scene support or interaction leaves the problem underconstrained, and without an admitted relation we use each participant's own evidence. Soft penetration and contact losses do not guarantee physically valid interaction in every frame.

The CMU evaluation covers Panoptic toddler interactions rather than the full range of child ages and environments.