Cross-Model Distillation of a Human-Pose Foundation Model from Unannotated Infant Video for Markerless Biomechanical Pose Estimation

R. James Cotton[1,2], Divya Joshi[1,3], and Colleen Peyton[3,4]

[1] Shirley Ryan AbilityLab, Chicago, IL, USA
[2] Department of Physical Medicine and Rehabilitation, Northwestern University, Chicago, IL, USA
[3] Department of Physical Therapy and Human Movement Science, Northwestern University, Chicago, IL, USA
[4] Department of Pediatrics, Northwestern University, Chicago, IL, USA

Abstract. Spontaneous movement is one of the earliest windows onto an infant's neuromotor health, and structured clinical instruments that score it are validated early predictors of cerebral-palsy risk. However, they require specially trained raters, are time-consuming, and carry inter-rater variability. This motivates automated, video-based markerless assessment, especially as marker-based motion capture is impractical in infants. Yet the foundation models that make markerless capture possible are trained almost entirely on adults: our recent multi-view infant study found that no single model is jointly best, with strong 2D keypoint accuracy and direct 3D body recovery split across different models [18]. While that study identifies this trade-off, it does not resolve it. Here, we perform cross-model distillation from the Sapiens 2 pose model into the SAM 3D Body model, using unannotated infant video alone. A frozen teacher supplies dense pseudo-labels, and a differentiable renderer aligns the predicted mesh to them in the training loop. On eleven held-out infants (18 sessions, 173 recordings) under our prior study's multi-view protocol, fine-tuning improves same-view 2D keypoint agreement with the Sapiens reference (median body percentage of correct keypoints @ 10px 0.22→0.42, face 0.22→0.42) and Procrustes-aligned mean per joint 3D position error (25.5→22.2 mm). This demonstrates how cross-model distillation improves SAM 3D Body model performance on infants.

Keywords: infant movement, markerless motion capture, pose estimation, foundation models, cross-model distillation, annotation-free adaptation, biomechanics, cerebral palsy

1 Introduction

Movement in early infancy is a direct window onto the developing nervous system, and the quality of an infant's spontaneous movement is among the earliest

observable signs of motor impairment [15, 24]. Clinicians exploit this through structured observation using validated instruments, such as the Prechtl General Movements Assessment [2, 12], the Test of Infant Motor Performance [32], and the Baby Observational Selective Control AppRaisal [1, 26, 27, 33], to predict later motor outcomes including cerebral palsy. However, these observational assessments depend on specially trained raters, are time-consuming, and carry inter-rater variability that limits their reach in routine care [2, 18, 23]. This motivates automated, video-based assessment that quantifies infant movement at scale. Markerless motion capture meets this need: it is non-invasive, scalable, and feasible in infants, a population in which affixing markers is impractical and often contraindicated [18]. This is already clinically actionable: kinematic features computed from 2D keypoints in single-camera infant video predict clinician-assessed General Movements Assessment scores in a large preregistered cohort of at-risk infants, using keypoint detectors never optimized for infants [31].

The clinically meaningful signal is not sparse 2D keypoints, but rather dense, whole-body 3D kinematics: the joint angles that describe how an infant actually moves, and whether individual joints move independently of one another — the selective motor control [6] whose emergence in infancy predicts later cerebral palsy [27]. Recovering these motion quantities from raw video pixels requires a 3D body model [8]. Building on the SMPL-style class of surface mesh models for humans [22, 25], we use the Momentum Human Rig (MHR) [13], which decouples a posable skeleton from a shape-dependent surface. Foundation models now regress these parameters directly from a single image: SAM 3D Body [35] predicts MHR parameters and in our prior study showed promising performance for infants [18]. However, models such as Sapiens and Sapiens 2 [19, 20] set the state of the art for 2D keypoint accuracy while not estimating joint angles.

These models are trained primarily on adults, who appear different visually, pose differently, and have different body proportions [3, 17, 30]. Furthermore, labeled infant pose data is scarce [14], and marker-based ground truth is infeasible [18]. Infant-specific shape spaces exist, such as the Skinned Multi-Infant Linear Model (SMIL) [16] and the lifespan model Anny [5], but lack foundation models for predicting those parameters from images.

Our recent work benchmarked MeTRAbs-ACAE [29], SAM 3D Body [35], and Sapiens [19, 20] on multi-view infant recordings [18]. Sapiens achieved the best 2D multi-view consistency but produced no direct 3D body estimates; SAM 3D Body produced fairly good 3D kinematics, but its predicted keypoints were not as geometrically consistent as those of Sapiens.

Here, we seek to close this gap by distilling pseudo-labels from a strong Sapiens 2 teacher into the SAM 3D Body model to improve predictions on infants, using only unlabeled infant video. A frozen Sapiens 2 teacher supplies dense pseudo-labels of keypoints, depth, surface normals, and silhouettes, and a differentiable renderer aligns the student's predicted mesh to them in the loop. On held-out infants this improves both 2D agreement and Procrustes-aligned 3D error.

2 Methods

Overview. We adapt a strong-3D human-pose foundation model to infants by distilling a strong-2D foundation-model teacher into it, using only unannotated infant video. The frozen teacher produces dense per-frame pseudo-labels, and a small trainable pose head is updated so that a differentiable render of the predicted body matches the teacher's pseudo-labels. We then re-run our established multi-view infant validation protocol [18] to quantify the gain (Figure 1).

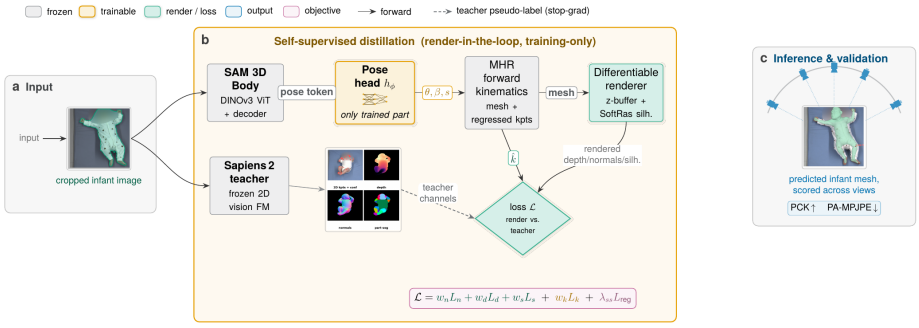

Fig. 1: Method overview. **(a)** A per-frame infant crop is the shared input to both teacher and student. **(b)** Render-in-the-loop cross-model distillation (training only): the frozen SAM 3D Body backbone and decoder feed a small pose head h_ϕ — the only trained part — whose MHR parameters drive differentiable forward kinematics and a differentiable renderer. A frozen Sapiens 2 teacher labels the same crop with dense pseudo-labels (2D keypoints, depth, surface normals, part segmentation); the loss $\mathcal{L}$ compares the render against those channels, backpropagating into the pose head alone. **(c)** At inference the adapted model runs per camera, validated against our prior study's multi-view benchmark.

2.1 Teacher, student, and the trainable surface

The teacher is Sapiens 2, a vision-transformer human-vision foundation model: it emits 2D keypoints with confidence scores, metric depth, surface normals, and body-part segmentation [20]. The infant crop and mask come from SAM 3 [7], prompted with the text "baby" so that the subject is chosen by concept rather than by size or position — the caregiver is often the larger and more central person in these recordings — and we keep only its highest-scoring detection. A frame is accepted only if that detection scores above 0.5, covers at least 5000 pixels, and does not run off the frame edge; after the teacher has run, the frame is also dropped unless the infant mask agrees with the teacher's own body segmentation, at an intersection-over-union above 0.35 or with at least 70% of the infant mask falling inside it. The teacher runs live during training, labelling frames on the fly. Its 2D keypoints follow the Goliath-308 layout: the first 70 are body and hand joints, and the remainder are dense facial landmarks [19, 20]. Distillation is supervised by these keypoints and their confidence scores, together with the

depth and normal maps. The body-part segmentation and the separate infant mask serve a gating role, controlling which pixels and keypoints each loss term is allowed to see.

The student is SAM 3D Body [35], which recovers a Momentum Human Rig (MHR) body [13] from a single top-down crop. A DINOv3 vision-transformer backbone encodes the crop, a promptable transformer decoder condenses it into a pose token, and a small feed-forward pose head projects that token to the MHR parameter vector: global orientation, body pose, per-hand PCA coefficients, facial expression, and identity shape (45-dim) and skeletal scale (28-dim) coefficients. Differentiable MHR forward kinematics — which decouples kinematic-tree scaling from pose — then maps these parameters to a posed body mesh and its regressed keypoints, which are the two outputs the distillation objective supervises. The mesh is passed to the renderer to produce depth, normal, and silhouette maps; the regressed keypoints follow the Goliath-308 layout (the first 70 are MHR body-and-hand joints, the remainder dense facial landmarks), matching the teacher so that all 308 keypoints can be supervised directly.

Distillation updates only the pose head's projection network — a two-layer feed-forward block of a few hundred thousand parameters — while the backbone, decoder, and MHR forward kinematics stay fixed. Within the pose head, the gradient is masked on the rows of the final linear layer that emit the 45 shape and 28 scale coefficients, so those output weights stay at their pretrained values while the rest of the head trains. The predicted shape and scale are not thereby held constant, since the features feeding those rows still adapt; what is removed is the direct route from the pixel losses into the shape readout. In the accompanying Supplementary Material, those rows are instead trained under an L2 prior λ_{ss}, swept in strength, and the comparison motivating the masked configuration is reported. The entire render-in-the-loop pipeline — teacher, student, renderer, and forward kinematics — is differentiable end-to-end on a MuJoCo-compatible JAX/Equinox stack [4, 21, 34].

2.2 Render-in-the-loop objective

For each accepted frame the student's pose head predicts the posed body from a fixed top-down crop. The predicted mesh is rendered differentiably and compared against the teacher's channels. Depth and surface normals are rendered by barycentric interpolation of the mesh under a z-buffered rasterizer, so gradients from those losses flow back through the render into the pose head. The silhouette is rendered with a soft rasterizer: per-pixel coverage is the largest of Gaussian kernels centred on the projected visible vertices, so it varies smoothly with vertex position and the silhouette loss is differentiable with respect to the outline. Back-face and depth culling remain a hard, detached test, so only the coverage field carries gradient. Per-vertex normals are mapped into the teacher's convention before comparison. These rendered outputs are compared against the teacher pseudo-labels through a combined objective,

$$\mathcal{L} = w_{\text{kpt}}\, L_{\text{kpt}} + w_{\text{normal}}\, L_{\text{normal}} + w_{\text{depth}}\, L_{\text{depth}} + w_{\text{silh}}\, L_{\text{silh}} + \lambda_{ss}\, L_{\text{reg}}, \quad (1)$$

in which the keypoint, normal, and depth terms are active and the silhouette term is weighted lightly.

The keypoint term L_{kpt} is a confidence-weighted robust loss on the 2D keypoints that separates major body joints, head anchors, and dense face landmarks, weighting the major joints and head anchors fully but the dense face landmarks an order of magnitude less. The teacher supplies 238 dense face landmarks against 12 major body joints and 5 coarse head anchors (nose, eyes and ears), so at equal weight the face — which is not the movement we set out to recover — would dominate the residual; down-weighting it produces clean, baby-sized fits. Teacher keypoints are also gated: those below the confidence threshold are dropped, and each remaining keypoint is looked up in the infant mask at its own pixel and given zero weight if it lands outside. This gating matters most for the caregiver, who is frequently in frame or in contact with the infant, so a hand or arm the teacher detects on the caregiver is excluded rather than fit, along with keypoints that fire on blankets and background clutter. The normal term L_{normal} is a cosine distance between rendered and teacher surface normals over an eroded body mask. Because the teacher predicts only relative monocular depth, the depth term L_{depth} is a per-frame affine-aligned L1: a closed-form scale-and-offset fit reconciles the two depth conventions and absorbs the teacher-versus-MHR camera-frame offset before comparison. The silhouette term L_{silh} aligns the rendered mesh boundary to the infant mask bidirectionally, retracting it where the mask excludes the body and extending it where the mask includes it. The Sapiens face/neck and hair segmentation classes are removed from the normal and depth losses and the silhouette term is disabled inside that region, since the adult-mean blendshapes cannot reach infant head proportions and pressure from the dense losses there deforms the body rather than the head. The regularizer L_{reg} is an L2 prior on shape and scale, inactive in the reported model because its shape and scale rows are masked; the sweep in the accompanying Supplementary Material document unfreezes them and varies λ_{ss}. All terms are differentiable back through the MHR forward kinematics.

2.3 Online data pipeline

The loader interleaves many recordings at once, so consecutive samples come from different infants, and within a recording it samples at a fixed stride of ten frames in short temporally-adjacent runs rather than independently. A cap on how many frames a recording may contribute before it is set aside forces a run to cover the cohort rather than the handful of recordings it happens to open with. A held-out tenth of the corpus, split by subject so that all timepoints of an infant fall on the same side, supplies an in-loop validation loss; every reported accuracy number comes from the separate multi-camera cohort described below.

2.4 Infant video corpus

The training corpus is a large set of raw, unannotated single-camera infant recordings. The reported cohort comprises 543 recordings of 311 unique infants

whose shorter side is at least 720 px, skewing preterm (median gestational age 31 weeks, IQR 27–38; median post-menstrual age at recording 52 weeks) — Table 1 summarizes the cohort demographics. The eleven validation infants are not part of this monocular corpus; they are additionally excluded at index build, so the cohort is leakage-free.

The eleven validation infants were recorded on the multi-camera system of our prior study [18]: eight synchronized RGB cameras (FLIR BlackFly S, f/1.4 lenses) at 29 frames per second, on tripods surrounding a mat with the infant supine at its centre, each session comprising 8–10 trials of roughly a minute of spontaneous movement. Intrinsics and extrinsics are calibrated per session from a moving checkerboard or ChArUco target, and the 3D reference is a robust multi-view triangulation that down-weights geometrically inconsistent detections [28] and low-confidence ones [9].

The parents and/or guardians of all participants provided informed, written consent to participate in this study, which was approved by the institutional review board of Northwestern University.

Table 1: Training-corpus demographics — the unlabeled monocular distillation cohort (543 recordings of 311 infants). Post-menstrual age is per recording; the other rows are per infant. The diagnosis row groups the clinical labels into *likely CP* (confirmed CP, 129, plus likely CP, 15) versus *unlikely CP* (typically developing, 157, plus unlikely CP, 8). Parenthetical counts give the denominator where a field was not recorded for every infant.

Characteristic	Value
Infants (unique)	311
Recordings	543
Recordings per infant, median [range]	1 [1–13]
Post-menstrual age at recording (wk), median [IQR], range	52 [42–54], 28–61
Gestational age (wk), median [IQR], range	31 [27–38], 22–53 (n = 305)
Preterm : term	204 : 105 (2 unrecorded)
Sex (M : F)	174 : 132 (5 unrecorded)
Diagnosis (likely CP : unlikely CP)	144 : 165 (2 unrecorded)

2.5 Implementation details

Only the pose head weights are optimized — excluding the masked shape and scale readout rows — using Adam with global-norm gradient clipping (clip 1.0, no weight decay). On a 40 GB GPU this system could only fit a batch size of two. The reported objective weights are $(w_{\text{kpt}}, w_{\text{normal}}, w_{\text{depth}}, w_{\text{silh}}) = (0.5, 0.5, 0.5, 0.1)$ with $\lambda_{ss} = 0$; keypoint sub-weights are 1.0 for major joints and head anchors and 0.1 for dense face landmarks, gated at teacher confidence 0.3. We train on the cohort above with the per-video coverage quota, photometric and crop-scale augmentation, and a peak learning rate of 2×10^{-5} reached by a short linear warmup and then held constant, for roughly 20k steps.

2.6 Evaluation protocol

We evaluate on that held-out multi-camera cohort using two readouts.

The first is per-camera 2D agreement: the student's 2D keypoints are scored against the Sapiens 2D keypoints in the *same* camera view, isolating the distillation gain from a monocular perspective. We report the percentage of correct keypoints (PCK) at fixed pixel radii (2, 5, and 10 px) over the body, face, and all-keypoint groups, under the same teacher-confidence gate used in training. Same-view metrics are computed on the earliest recording of each infant so that every infant contributes equally.

The second is PA-MPJPE: the student's monocular 3D keypoints are compared against a multi-view triangulated [28] reference under a per-frame Procrustes alignment with scale, making the metric invariant to the global similarity transform a monocular setup cannot recover. The alignment is solved over eleven reliably triangulated body joints (shoulders, elbows, wrists, knees, ankles, and neck as root); we additionally report a thirteen-joint variant that includes the hips, which are otherwise excluded due to poor triangulation in occluded infants. Both the 2D same-view targets and the 3D triangulated reference are derived from Sapiens 2D estimates across the calibrated multi-camera rig, following our prior study's protocol [18] — they are model-derived references rather than marker-based ground truth, so both readouts measure agreement with a strong automatic reference, not an independent gold standard.

All frames are sampled at a fixed stride of ten before scoring. We report the median across infants with a normalized inter-quartile range ($\sigma_{\mathrm{NIQR}} = 0.7413 \cdot \mathrm{IQR}$) as an outlier-robust spread, with mean and standard deviation reported alongside, and assess significance with a paired Wilcoxon signed-rank test across the eleven infants.

Both readouts are reported in Table 2 alongside our prior study's multi-view benchmark harness — reprojection error, geometric consistency, and a broader-keypoint PA-MPJPE — rerun on the same held-out infants for direct comparison [18].

Geometric consistency GC_d [9] is the fraction of detected 2D keypoints that a reconstruction reprojects to within d pixels, counting only detections whose confidence exceeds λ. Writing $\delta_{t,j,c} = \|\Pi_c \mathbf{x}_{t,j} - y_{t,j,c}\|$ for the image-plane distance in camera c between the triangulated 3D keypoint $\mathbf{x}_{t,j}$ reprojected through that camera's projection operator Π_c (including intrinsic distortion) and the corresponding 2D detection $y_{t,j,c}$ of confidence $w_{t,j,c}$,

$$\mathrm{GC}_d = \frac{\sum_{t,j,c} \mathbb{1}(\delta_{t,j,c} < d)\, \mathbb{1}(w_{t,j,c} > \lambda)}{\sum_{t,j,c} \mathbb{1}(w_{t,j,c} > \lambda)}, \tag{2}$$

with $\lambda = 0.5$ and, following our prior study, non-facial keypoints only. We report GC_5 and GC_{10}; higher is better.

3 Results

We evaluate on 11 held-out multi-camera infants (18 capture sessions, 173 recordings), entirely separate from the distillation corpus by construction, sampling

every recording at a fixed frame stride of ten. Table 2 consolidates all quantitative results into a single table, scored on the same strided frames of the same held-out infants, and groups metrics by their reference rather than by view protocol. The **multi-view consistency** rows replicate our previous paper's harness — reprojection error and geometric consistency (GC) — and carry the Sapiens teacher's published values for reference ([18]). The **teacher-agreement** rows are our own readout: per-camera 2D PCK@10px against the Sapiens teacher's keypoints in the same view, and Procrustes-aligned 3D error (PA-MPJPE) over reliably triangulated core body joints, reported as the median (nIQR) across infants. Procrustes-aligned 3D error is reported over three nested keypoint sets. The first two are scored against a Sapiens-triangulated reference and the third against a reference triangulated from the MHR keypoints themselves, so the third is not directly comparable with the first two. The first is eleven core body joints — shoulders, elbows, wrists, knees, ankles and neck — and is our primary 3D metric. The second adds both hips, which the reference localizes poorly. The third is MHR-70, the model's complete keypoint output; 40 of those 70 are finger keypoints. The first two solve the alignment on the joints they score, the third over the full 70-keypoint layout.

3.1 Qualitative results

Figure 2 compares the recovered MHR mesh between the base SAM-3D-Body model and the fine-tuned student on four example frames from held-out infants, viewed from the rig's most elevated camera, which looks down on the mat. Fine-tuning consistently reduces the 2D distance to the Sapiens teacher keypoints, though the fitted mesh also exhibits shape artifacts — an adult-proportioned face and an overly narrow, developed torso — that persist after distillation.

Figure 3 compares the Sapiens pseudo-label channels against the student's rendered predictions. The shape artifacts visible in Figure 2 are absent in the pseudo-labels themselves.

3.2 Quantitative results

Distillation moves the student toward the Sapiens teacher (Table 2); teacher-agreement metrics are reported as the median across infants with the normalized inter-quartile range ($\sigma_{\mathrm{NIQR}} = 0.7413 \cdot \mathrm{IQR}$) as an outlier-robust spread. Keypoint agreement (PCK@10px, model vs. teacher in the same camera) approximately doubles — body 0.216 $\rightarrow$0.418 and face 0.219 $\rightarrow$0.422 — and the scale-invariant 3D PA-MPJPE over the reliably triangulated core body joints improves from 25.5 $\rightarrow$22.2 mm, all significant under a paired Wilcoxon signed-rank test ($p \leq 0.005$). Pixel error falls correspondingly (Figure 4).

We report PA-MPJPE primarily over the eleven body joints that exclude the hips, because the triangulated reference localizes infant hips unreliably due to frequent occlusions. Including the hips leaves the improvement direction intact (median 33.4 $\rightarrow$25.1 mm) but sharply inflates the spread ($\sigma_{\mathrm{NIQR}} \approx 20$ mm),

Fig. 2: Qualitative comparison on four held-out infants (rows), through the capture rig's shared elevated camera. Columns: face-blurred input; base SAM-3D-Body mesh; fine-tuned mesh. For each row we highlight the major limb joint where fine-tuning most reduced the 2D distance to the Sapiens teacher keypoint for that frame and camera; a black star marks the Sapiens target, and the line and label on each mesh column give that model's pixel distance to it (fine-tuned consistently closer). Targets are the Sapiens-Goliath 2D keypoints, computed as in our prior study.

driven by a few recordings whose reference hips fail. Figure 4 breaks these improvements out per infant: gains hold across almost the entire cohort.

On our prior study's benchmark, fine-tuning improves over the SAM-3D-Body base on reprojection error (39.8 →36.7 px) and geometric consistency (GC@10 0.31 →0.36). It remains well below off-the-shelf Sapiens on 2D metrics (reprojection 22.8 px, GC@10 0.82), as expected given that Sapiens is the dis-

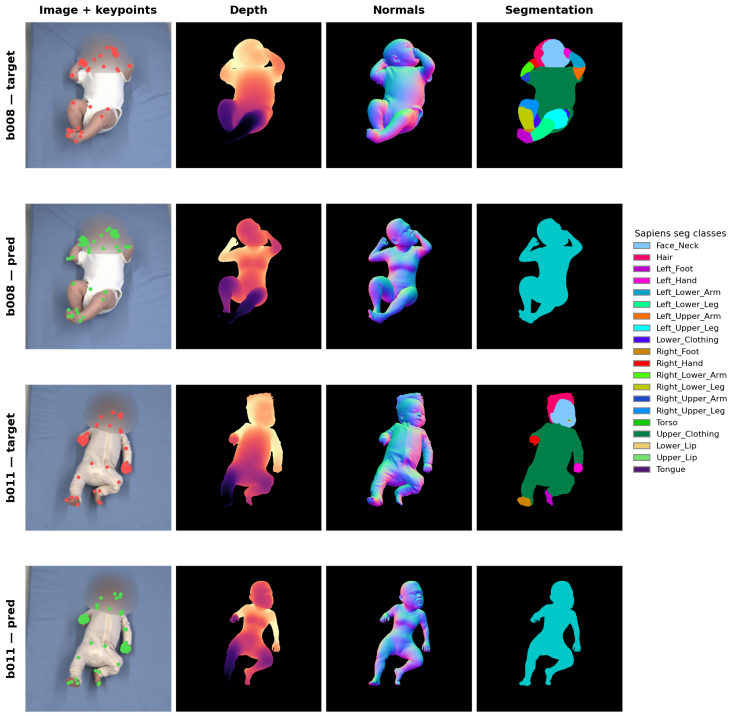

Fig. 3: Sapiens teacher pseudo-labels (target rows) vs fine-tuned MHR predictions (pred rows) for two infants. Columns: 2D keypoints on the face-blurred photo, depth, normals, segmentation. The segmentation channel was not used directly in supervision; only the binary infant mask was.

tillation teacher. The Sapiens column carries our prior study's published values ([18]); the SAM-3D-Body and Ours columns are computed through our pipeline over the same 11 infants.

3.3 Shape–scale regularization

The reported model masks the gradient on the shape and scale rows of the pose head's output layer. Training those rows instead, under an L2 prior λ_{ss} swept from strong (0.05) through light (0.005) to none (0), does not improve accuracy: on the held-out reference the masked model is best or tied for best on every metric, and with no prior ($\lambda_{ss} = 0$) body proportions inflate, most visibly the head, into shapes atypical of infants. The full sweep — a combined accuracy table over all four settings at their final checkpoints together with the corresponding mesh overlays — is reported in the accompanying Supplementary Material document (Table S1, Fig. S1), submitted alongside this paper.

Table 2: Accuracy on 11 held-out infants. *Multi-view consistency* rows are means over infants; *teacher agreement* rows are medians (σ_{NIQR}). The three PA-MPJPE rows are nested keypoint sets, each Procrustes-aligned per frame with scale on the keypoints it scores: **11 core joints** (shoulders, elbows, wrists, knees, ankles, neck), our primary 3D metric; **+ hips**, whose reference positions are less reliable in occluded infants; and **MHR-70**, the model's full output, of which 40 keypoints are fingers. MHR-70 is scored against a different triangulated reference from the other two and is not comparable with them. Sapiens produces no monocular 3D body (—). PCK and GC are higher-is-better, pixel error and PA-MPJPE (mm) lower-is-better; best value per row in **bold**.

Metric	Sapiens	SAM-3D-Body	**Ours**
Multi-view consistency			
Reprojection (px) ↓	**22.8**	39.8	36.7
GC@10 ↑	**0.82**	0.31	0.36
Teacher agreement			
Body PCK@10 ↑	—	0.216 (0.08)	**0.418 (0.09)**
Face PCK@10 ↑	—	0.219 (0.03)	**0.422 (0.09)**
PA-MPJPE, 11 core joints (mm) ↓	—	25.5 (5.4)	**22.2 (3.6)**
PA-MPJPE, + hips (13 kpt) (mm) ↓	—	33.4 (20.8)	**25.1 (20.1)**
PA-MPJPE, MHR-70 (mm) ↓	—	27.3 (10.1)	**23.7 (7.1)**

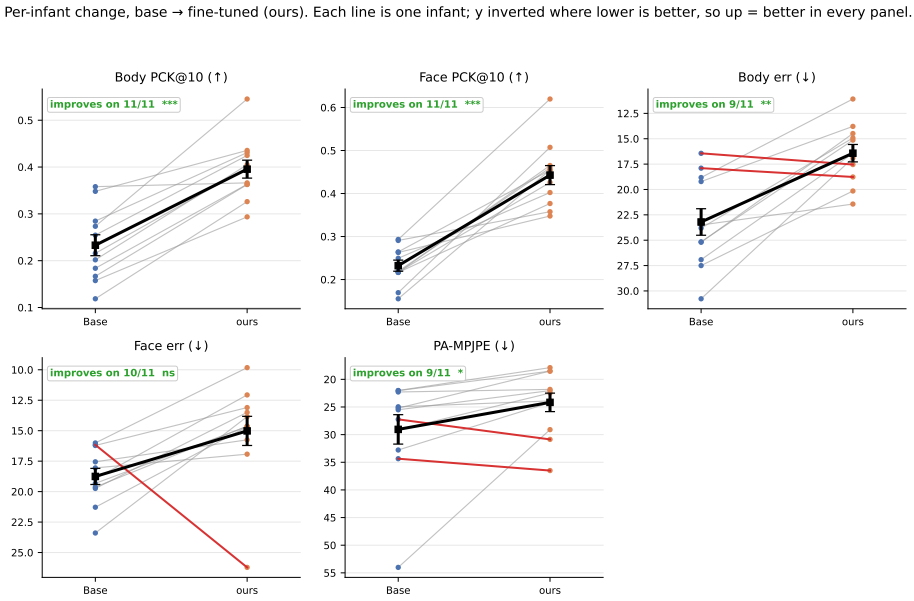

Fig. 4: Per-infant change, base →fine-tuned, one line per infant with the heavy black line the cohort mean (± 1 SEM). Panels for lower-is-better metrics have the y-axis inverted so "up = better" reads identically everywhere; regressing infants, where any, are drawn in red — the two PCK panels improve on all 11 infants, and the pixel-error and PA-MPJPE panels have one to two regressions each.

3.4 Training loss composition

The objective is dominated by the 2D keypoint terms; the render-based dense terms (surface normals, silhouette, depth) contribute little and do not converge. Table 3 decomposes the converged in-loop objective into each term's weighted contribution (weight × loss value) as a fraction of the total.

Table 3: Loss-term composition of the reported run (converged in-loop eval values; the weighted terms sum to the total loss). The keypoint terms dominate the objective; the dense render terms are a small, non-converging fraction.

Term	Value	Weight	Weighted	%
L_{kpt}, major joints	2.37	1.0	2.37	32.6
L_{kpt}, head	2.40	1.0	2.40	33.0
L_{kpt}, body	3.43	0.5	1.72	23.6
L_{kpt}, face	6.57	0.1	0.66	9.0
L_{normal}	0.185	0.5	0.093	1.3
L_{silh}	0.243	0.1	0.024	0.3
L_{depth}	0.037	0.5	0.019	0.3

The keypoint terms together account for 98.1% of the objective; the dense terms (normals + silhouette + depth) sum to 1.86%.

4 Discussion

Our prior study [18] demonstrated that two recent foundation models for human pose estimation exhibit complementary strengths: SAM 3D Body produces joint angles for an MHR kinematic model with reasonable 3D accuracy, while Sapiens achieves more precise 2D joint position estimates but lacks 3D joint angle information. This work shows that these complementary strengths can be combined through label-free cross-model distillation. We distill the 2D fidelity of Sapiens 2 into SAM 3D Body through a render-in-the-loop objective, using only unannotated infant video. On eleven held-out infants with multi-view data to provide high quality reference measurements (173 recordings), the fine-tuned student shows statistically significant improvements in 2D agreement with the Sapiens reference along with scale-invariant 3D error (PA-MPJPE), narrowing the trade-off our prior study identified without requiring any manual annotation or marker-based ground truth.

Recovering whole-body 3D kinematics, rather than 2D keypoints alone, from ordinary video is a meaningful step toward scalable video-only assessment of early motor development, particularly in settings where marker-based capture is impractical and contraindicated [18]. The quality of selective motor control — the ability to move individual joints independently of one another — is a clinically important indicator of motor impairment, and its emergence in infancy predicts later cerebral palsy [6, 27]. While the absolute improvement in 3D accuracy is a few millimetres, relative to the size of an infant this is about a 10% reduction in Procrustes-aligned error (25.5 to 22.2 mm). Our goal is ultimately to classify movements, reproduce clinical scores, and find clinically meaningful

patterns in movement; it remains to be seen what level of accuracy is required for accurate classification of pathological movements. Notably, kinematic features computed from 2D keypoints alone, using a pose estimator never tuned for infants, already predict clinician-assessed General Movements Assessment scores [31]. Recovering 3D biomechanical kinematics should nonetheless provide a viewpoint-invariant and anatomically interpretable representation on which such classifiers can be built.

A parallel line of our work concerns what the model outputs rather than how well it tracks. Mesh recovery returns vertices and keypoints posed on non-anatomical kinematic trees and not the biomechanical joint angles. In concurrent work we added a biomechanical head to the same SAM 3D Body backbone that regresses the joint angles and segment scales of a biomechanical model directly from one image [11]. That head is supervised by in-loop optimized targets from an inverse-kinematics solver we developed for monocular finger tracking [10] rather than by paired biomechanical labels, so like the distillation reported here it trains on unlabeled images. Combining the two is a natural next step: the distillation described in this paper adapts the mesh to infants, and biomechanical distillation on top of it would return infant movement as joint angles from which clinical measures can be computed directly.

This work has several limitations. The most pronounced is that while the keypoint accuracy improves, both in 2D and 3D, the shape of the mesh still differs from the infant's, particularly at the head. This may indicate the need to further tune the hyperparameters of our distillation. The non-keypoint-based losses contributed only 2% to the final loss at convergence, despite Figure 3 showing substantial errors in the silhouette. MHR's shape space was designed for adults and may not well represent the large head-to-body ratio, short limbs, and distinct limb-to-torso proportions of early development [17, 30], so even a well-distilled student is fitting baby movement onto an adult-shaped body. This adult-mean limitation motivates a different shape model. ANNY [5] is a natural candidate: a scan-free, fully differentiable parametric model whose interpretable age phenotype spans infants through elders. No foundation model yet predicts ANNY parameters with infant surface contours, so developing one is a natural future direction alongside further hyperparameter tuning.

While we test on a held-out dataset, our videos were largely obtained in a format designed for assessing motor performance, with infants supine and alone, and with unsettled movement minimized. Because of this, generalization to other contexts, such as crawling or being held, remains to be seen. A more fundamental limitation is the absence of ground-truth annotations. Manual keypoint labeling and physical marker-based motion capture, the gold standards for pose estimation evaluation, are both highly impractical and largely contraindicated in this population: affixing markers to preterm or clinically monitored infants is invasive, disruptive, and in many settings not permitted. We therefore follow our prior study [18] in using multi-view triangulation as the next best alternative, which provides a high-quality reference for 3D accuracy [28]. However, triangulation is itself model-derived and carries its own error, particularly at frequently

occluded joints such as the hips, so our accuracy estimates are relative to an imperfect reference rather than an absolute ground truth.

4.1 Conclusion

Taken together, these results show that adapting a 3D body model to a data-scarce infant population does not require infant pose labels or marker-based ground truth — only a frozen teacher, unlabeled video, and a differentiable render-in-the-loop objective to connect the two. This work moves toward markerless tools that could one day support early motor assessment at scale from monocular video, in exactly the settings where such tools are needed most.

Acknowledgements

We would like to thank Imani Mann and Grace Hoo for their assistance with participant recruitment and scheduling, Shawana Anarwala and Kayan Abdou for their assistance with experimental setup and data collection, and Kunal Shah for his assistance with infrastructure.

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number R21HD117074, and by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number T32TR005124. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Barbosa, V.M., Peyton, C., Sukal-Moulton, T.: Baby Observational Selective Control AppRaisal (BabyOSCAR): Construct validity and test performance. Developmental Medicine and Child Neurology 66(11), 1502–1510 (2024)

[2] Bosanquet, M., Copeland, L., Ware, R., Boyd, R.: A systematic review of tests to predict cerebral palsy in young children. Developmental Medicine and Child Neurology 55(5), 418–426 (2013)

[3] Bose, S., Dela Cruz, H., Dutta, A., Kokkoni, E., Karydis, K., Roy-Chowdhury, A.K.: Leveraging Synthetic Adult Datasets for Unsupervised Infant Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5601–5610 (2025)

[4] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: Composable Transformations of Python+NumPy Programs. https://github.com/google/jax (2018), https://github.com/google/jax

[5] Brégier, R., Fiche, G., Bravo-Sánchez, L., Lucas, T., Armando, M., Weinzaepfel, P., Rogez, G., Baradel, F.: Human Mesh Modeling for Anny Body. https://arxiv.org/abs/2511.03589 (2025), https://arxiv.org/abs/2511.03589, arXiv:2511.03589

[6] Cahill-Rowley, K., Rose, J.: Etiology of impaired selective motor control: emerging evidence and its implications for research and treatment in cerebral palsy. Developmental Medicine and Child Neurology 56(6), 522–528 (6 2014)

[7] Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Coll-Vinent, D.S., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollar, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment Anything with Concepts. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=r35clVtGzw

[8] Cotton, R.J.: Differentiable Biomechanics Unlocks Opportunities for Markerless Motion Capture. In: 2025 International Conference On Rehabilitation Robotics (ICORR). pp. 44–51 (2025)

[9] Cotton, R.J., Cimorelli, A., Shah, K., Anarwala, S., Uhlrich, S., Karakostas, T.: Improved Trajectory Reconstruction for Markerless Pose Estimation. In: 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2023), https://ieeexplore.ieee.org/document/10340745, defines the GC@Npx geometric-consistency metric.

[10] Cotton, R.J., Firouzabadi, P., Murray, W.M.: Monocular Biomechanical Tracking of Fingers with Inverse Kinematics to Foundation Models. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2026)

[11] Cotton, R.J., Peiffer, J.D., Williamson, L., Leske, J., Pavlakos, G.: Biomechanical 3D Body: Self-Supervised Distillation of Biomechanical Pose from a 3D Body Foundation Model. In: ECCV Workshop on Human Motion Challenges in Real-World and Clinical Settings (MoCha) (2026)

[12] Einspieler, C., Prechtl, H.F.R.: Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system. Mental Retardation and Developmental Disabilities Research Reviews 11(1), 61–67 (2005)

[13] Ferguson, A., Osman, A.A.A., Bescos, B., Stoll, C., Twigg, C., Lassner, C., others: MHR: Momentum Human Rig (2025), arXiv:2511.15586

[14] Gama, F., Míšař, M., Navara, L., Popescu, S.T., Hoffmann, M.: Automatic infant 2D pose estimation from videos: Comparing seven deep neural network methods. Behavior Research Methods 57(10), 280 (2025)

[15] Herskind, A., Greisen, G., Nielsen, J.B.: Early identification and intervention in cerebral palsy. Developmental Medicine and Child Neurology 57(1), 29–36 (2015)

[16] Hesse, N., Pujades, S., Black, M.J., Arens, M., Hofmann, U.G., Schroeder, A.S.: Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D Sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 42(10), 2540–2551 (2020), https://arxiv.org/abs/1810.07538

[17] Huang, X., Fu, N., Liu, S., Ostadabbas, S.: Invariant Representation Learning for Infant Pose Estimation with Small Data. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–8. IEEE (2021)

[18] Joshi, D., Peiffer, J.D., Peyton, C., Cotton, R.J.: Markerless Motion Capture for Biomechanical Whole-Body Kinematic Estimation in Infants. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2026), https://arxiv.org/abs/2605.17120

[19] Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for Human Vision Models. In: Computer Vision – ECCV 2024. Lecture Notes in Computer Science, vol. 15062, pp. 206–228. Springer (2024)

[20] Khirodkar, R., Wen, H., Martinez, J., Dong, Y., Zhaoen, S., Saito, S.: Sapiens2. In: International Conference on Learning Representations (ICLR) (2026), arXiv:2604.21681

[21] Kidger, P., Garcia, C.: Equinox: neural networks in JAX via callable PyTrees and filtered transformations. Differentiable Programming workshop at Neural Information Processing Systems 2021 (2021), https://arxiv.org/abs/2111.00254

[22] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. ACM Transactions on Graphics (Proc. SIGGRAPH Asia) 34(6), 248:1–248:16 (2015)

[23] Noble, Y., Boyd, R.: Neonatal assessments for the preterm infant up to 4 months corrected age: A systematic review. Developmental Medicine and Child Neurology 54(2), 129–139 (2012)

[24] Novak, I., Morgan, C., Adde, L., Blackman, J., Boyd, R.N., Brunstrom-Hernandez, J., Cioni, G., Damiano, D., Darrah, J., Eliasson, A.C., de Vries, L.S., Einspieler, C., Fahey, M., Fehlings, D., Ferriero, D.M., Fetters, L., Fiori, S., Forssberg, H., Gordon, A.M., Greaves, S., Guzzetta, A., Hadders-Algra, M., Harbourne, R., Kakooza-Mwesige, A., Karlsson, P., Krumlinde-Sundholm, L., Latal, B., Loughran-Fowlds, A., Maitre, N., McIntyre, S., Noritz, G., Pennington, L., Romeo, D.M., Shepherd, R., Spittle, A.J., Thornton, M., Valentine, J., Walker, K., White, R., Badawi, N.: Early, accurate diagnosis and early intervention in cerebral palsy: Advances in diagnosis and treatment. JAMA Pediatrics 171(9), 897–907 (9 2017)

[25] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: 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 (CVPR). pp. 10975–10985 (2019), https://arxiv.org/abs/1904.05866

[26] Peyton, C., Aaby, D., Barbosa, V.M., Boswell, L., de Regnier, R.A., Bos, A.F., Sukal Moulton, T.: Baby Observational Selective Control AppRaisal (BabyOSCAR): Scores at 3 months predict functional ability, spastic cerebral palsy distribution, and diagnosis at 2 years. Developmental Medicine and Child Neurology 66(11), 1521–1528 (2024)

[27] Peyton, C., Moulton, T.S., Aaby, D., Millman, R., deRegnier, R.A., Bos, A.F., Dewald, J.P.: Emergence of Selective Motor Control in Early Infancy: A Longitudinal Study of Infants Born Preterm with and without Cerebral Palsy. The Journal of Pediatrics 289, 114915 (2026), https://www.sciencedirect.com/science/article/pii/S0022347625004561

[28] Roy, S.K., Citraro, L., Honari, S., Fua, P.: On Triangulation as a Form of Self-Supervision for 3D Human Pose Estimation. In: Proceedings - 2022 International Conference on 3D Vision, 3DV 2022 (2022)

[29] Sárándi, I., Hermans, A., Leibe, B.: Learning 3D Human Pose Estimation from Dozens of Datasets using a Geometry-Aware Autoencoder to Bridge Between Skeleton Formats. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2956–2966 (2023), http://arxiv.org/abs/2212.14474

[30] Sciortino, G., Farinella, G.M., Battiato, S., Leo, M., Distante, C.: On the Estimation of Children's Poses. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) Image Analysis and Processing - ICIAP 2017. pp. 410–421. Springer International Publishing (2017)

[31] Segado, M., Prosser, L.A., Duncan, A.F., Johnson, M.J., Kording, K.P.: A preregistered, open pipeline for early cerebral palsy risk assessment from infant videos. GigaScience 15, giag003 (2026)

[32] Spittle, A.J., Doyle, L.W., Boyd, R.N.: A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life. Developmental Medicine and Child Neurology 50(4), 254–266 (2008)

[33] Sukal-Moulton, T., Barbosa, V.M., Sargent, B., Boswell, L., de Regnier, R.A., Bos, A.F., Peyton, C.: Baby Observational Selective Control AppRaisal (BabyOSCAR): Convergent and discriminant validity and reliability in infants with and without spastic cerebral palsy. Developmental Medicine and Child Neurology 66(11), 1511–1520 (2024)

[34] Todorov, E., Erez, T., Tassa, Y.: MuJoCo: A Physics Engine for Model-Based Control. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2012)

[35] Yang, X., Kukreja, D., Pinkus, D., Fan, T., Park, J., Shin, S., Cao, J., Liu, J.W., Ugrinovic, N., Sagar, A., Malik, J., Feiszli, M., Dollár, P., Kitani, K.: SAM 3D Body: Robust Full-Body Human Mesh Recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7209–7219 (2026), arXiv:2602.15989

Supplementary Material — Cross-Model Distillation of a Human-Pose Foundation Model from Unannotated Infant Video for Markerless Biomechanical Pose Estimation

R. James Cotton[1,2], Divya Joshi[1,3], and Colleen Peyton[3,4]

[1] Shirley Ryan AbilityLab, Chicago, IL, USA
[2] Department of Physical Medicine and Rehabilitation, Northwestern University, Chicago, IL, USA
[3] Department of Physical Therapy and Human Movement Science, Northwestern University, Chicago, IL, USA
[4] Department of Pediatrics, Northwestern University, Chicago, IL, USA

Abstract. This supplementary document reports the shape/scale regularization sweep referenced in the main paper: the reported model, whose shape and scale readout rows are gradient-masked, against three runs that train those rows under an L2 prior of varying strength, scored through the identical evaluation harness on the same eleven held-out infants.

1 Supplementary Material

1.1 Shape/scale constraint: does training the readout help?

The reported model masks the gradient on the shape and scale rows of the pose head's output layer (Methods). Here we ask whether *training* those rows — under an L2 prior of weight λ_{ss}, swept from strong through none — improves the fit. This spans a spectrum from the masked model, through progressively weaker priors, to no prior at all.

We compare four training runs — the masked reported model and three trained-readout runs at $\lambda_{ss} \in \{0.05, 0.005, 0\}$ — each evaluated at its **final** checkpoint (no loss-minimum selection), through the identical evaluation harness used in the main results, on the same eleven held-out infants. Table S1 reports the combined accuracy metrics and Figure S1 shows the corresponding meshes on held-out infants. Both are produced from the four final checkpoints by a single script that runs the evaluations, aggregates the table, and renders the overlay figure end to end.

This sweep is a negative result (Table S1): the masked model is best on every held-out metric — or tied for best, as it matches the light prior ($\lambda_{ss} = 0.005$) on geometric consistency — and training those rows does not improve accuracy. The appeal of training those rows is that the coefficients can deform

Table S1: Shape/scale-constraint ablation on the 11 held-out infants, each run evaluated at its FINAL checkpoint and ordered by decreasing shape/scale constraint (readout masked ->readout trained with a vanishing shape regularizer λ). The three trained-readout runs also carry rebalanced dense-loss weights (normals, depth and silhouette raised by roughly 26x, 128x and 100x relative to the masked run), so they vary the readout constraint and the dense weighting together rather than the constraint alone. Teacher-agreement (T) columns are the MEDIAN across infants of the per-camera 2D agreement with the Sapiens teacher (PCK@10px) and the hips-excluded core PA-MPJPE. Benchmark (B) columns are mean ± std over infants from the prior study's harness (non-facial reprojection error and GC@10; PA-MPJPE over the full MHR-70). PCK and GC are higher-is-better; pixel error and PA-MPJPE (mm) are lower-is-better.

Run	Constraint	λ	Body PCK@10 (T) ↑	Face PCK@10 (T) ↑	PA-MPJPE–hips (T, mm) ↓	Reproj (B, px) ↓	GC@10 (B) ↑	MHR-70 PA-MPJPE (B, mm) ↓
SAM-3D-Body base	n/a	–	0.216	0.219	25.5	39.8 +/- 5.7	0.31 +/- 0.12	32.4 +/- 13.7
Masked (reported model)	masked	n/a	0.418	0.422	22.2	36.7 +/- 3.8	0.36 +/- 0.10	28.8 +/- 11.4
Trained readout, no reg	trained	0	0.376	0.355	22.5	38.0 +/- 4.4	0.33 +/- 0.09	29.2 +/- 9.8
Trained readout, light reg	trained	0.005	0.392	0.409	23.2	37.4 +/- 4.7	0.36 +/- 0.11	29.0 +/- 10.8
Trained readout, strong reg	trained	0.05	0.342	0.227	29.4	40.7 +/- 5.2	0.26 +/- 0.08	43.2 +/- 13.3

the body toward whatever configuration best minimizes the pixel losses — and with no prior ($\lambda_{ss} = 0$) this does reach the tightest in-loop (teacher-fidelity) fit — but that freedom is spent deforming *shape* rather than improving 3D joint accuracy: the unregularized meshes inflate the head and torso into anatomically implausible shapes (Figure S1) without beating the masked model on the held-out reference. A strong prior ($\lambda_{ss} = 0.05$, trained twice as long) is worst of all; the light prior $\lambda_{ss} = 0.005$ is the best of the trained-readout runs — matching the masked model on geometric consistency but trailing it on every other metric. We therefore report the masked model. Its limitation is the counterpart to its strength: masking the readout keeps the meshes clean but still does not reach true infant-specific body shape — a limitation that joint-position metrics alone do not capture, and that motivates extending the method to an infant-aware shape basis such as Anny [1].

The in-loop loss measures fidelity to the teacher, not accuracy against the held-out reference. The unregularized run ($\lambda_{ss} = 0$) reaches the lowest in-loop loss of the three trained-readout runs and is still not the best of them on the held-out metrics, let alone competitive with the masked model we report. This

Fig. S1: Predicted MHR mesh overlaid on held-out infants across the shape/scale sweep: input, then the masked model, $\lambda_{ss} = 0$, $\lambda_{ss} = 0.005$, and $\lambda_{ss} = 0.05$. All meshes register to the same person under the same camera. The unregularized run ($\lambda_{ss} = 0$) produces a mixed set of shape changes: the torso becomes somewhat less narrow than in the reported model — moving in the direction of infant proportions — but the head simultaneously enlarges and rounds while the forehead narrows in an anatomically implausible way. The strongly regularized run and the reported model stay closer to the base model's proportions.

mirrors the teacher-fidelity/accuracy gap discussed in the main text: a lower training loss does not guarantee better held-out performance.

References

[1] Brégier, R., Fiche, G., Bravo-Sánchez, L., Lucas, T., Armando, M., Weinzaepfel, P., Rogez, G., Baradel, F.: Human Mesh Modeling for Anny Body. https://arxiv.org/abs/2511.03589 (2025), https://arxiv.org/abs/2511.03589, arXiv:2511.03589