Cross-Model Distillation of a Human-Pose Foundation Model from Unannotated Infant Video for Markerless 3D Pose Estimation

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

[1]Shirley Ryan AbilityLab [2]PM&R, Northwestern [3]Physical Therapy & Human Movement Sciences, Northwestern [4]Pediatrics, Northwestern

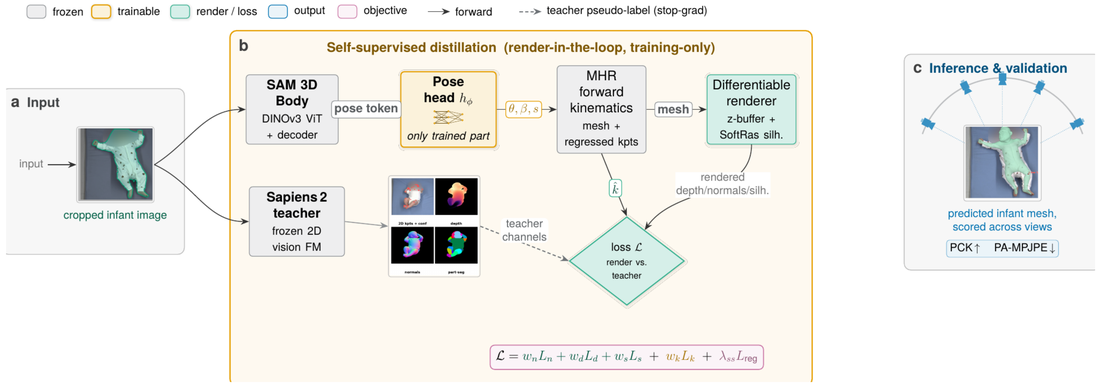
Overview

Problem Spontaneous movement is the earliest window onto an infant's neuromotor health, but the instruments that score it need trained raters. Markerless capture is the scalable alternative — yet pose foundation models are trained on adults, infant labels are scarce, and affixing markers is impractical in infants.

Approach **Cross-model distillation from unannotated infant video alone**: a frozen Sapiens 2 teacher supplies dense pseudo-labels (2D keypoints, depth, normals, segmentation) and a **differentiable renderer** aligns the SAM-3D-Body student's mesh to them in the training loop. Only the pose head moves.

Result On 11 held-out infants (18 sessions, 173 recordings) keypoint agreement with the teacher **roughly doubles** (body PCK@10 $0.22 \rightarrow 0.42$) and 3D PA-MPJPE improves $25.5 \rightarrow 22.2$ **mm** — with **no labels and no markers**.

Method — render-in-the-loop cross-model distillation



Teacher — frozen Sapiens 2

- Dense pseudo-labels, emitted *live*: 2D keypoints + confidences, depth, normals, segmentation.
- Crop & mask: SAM 3 prompted **"baby"** — selects the subject by concept, not size.

Student — SAM-3D-Body

- Only the **pose head** trains (backbone, decoder, FK frozen); gradient **masked on shape & scale rows** — unfreezing them inflates proportions.

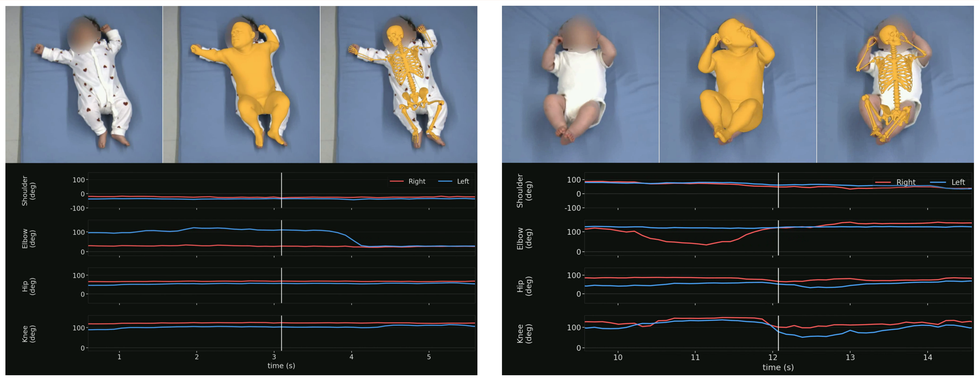
Corpus

- 543 unannotated recordings, 311 infants; skews preterm (median gestational age 31 wk).
- 11 validation infants excluded at index build — leakage-free split.

Objective

- $\mathcal{L} = w_{\text{kpt}} L_{\text{kpt}} + w_{\text{normal}} L_{\text{normal}} + w_{\text{depth}} L_{\text{depth}} + w_{\text{silh}} L_{\text{silh}}$
- Keypoints: confidence-weighted, *mask-gated* (caregiver limbs excluded, not fit) — **98%** of the converged objective.
- Depth affine-aligned; soft-rasterized silhouette gives outline gradient.

Since this paper — distilling infant *biomechanics*



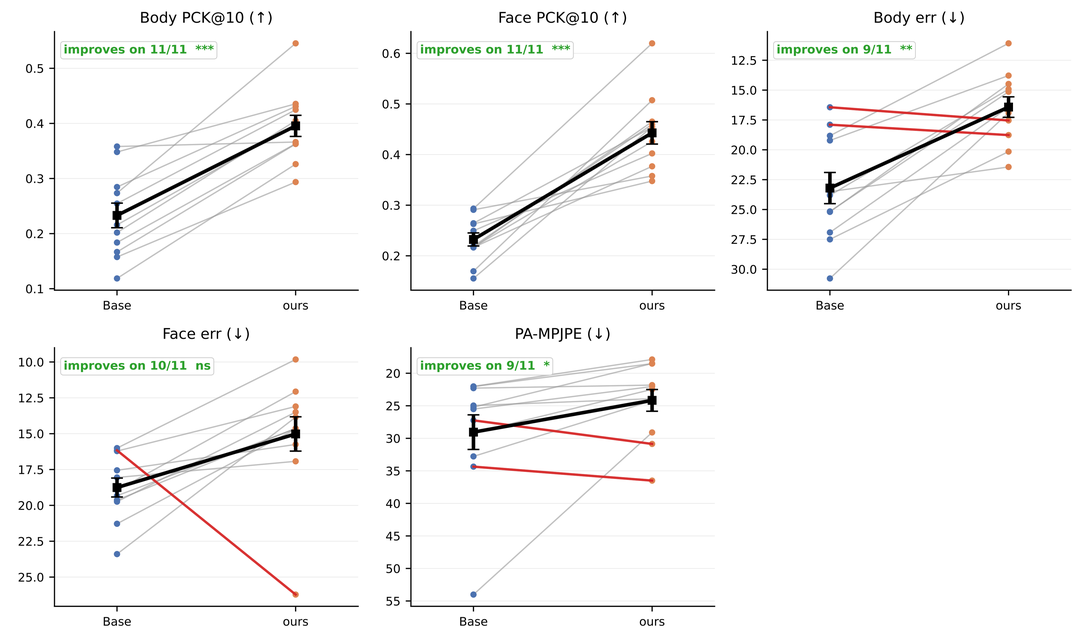
Two held-out infants: **tracked crop** | **MHR mesh** | **biomechanical skeleton** + joint angles.

- A companion **biomechanical head** now trains *together* with the distillation above — emitting infant movement as **joint angles**.
- Three epochs in, in-the-loop IK barely corrects it: median **hip 1.0–1.1°, knee 0.3–0.5°, elbow 0.6–0.7°** (from 5–15° and 5–11°); skeleton outside the mesh **6.9% → 1.3%**.
- **Preliminary:** the raw head is still $4\times$ its IK-refined output; one held-out monocular clip got slightly worse; **no infant biomechanical ground truth exists to score it against.**

Method: *Biomechanical 3D Body: Self-Supervised Distillation of Biomechanical Pose from a 3D Body Foundation Model*, Cotton et al., ECCV Workshop on Human Motion Challenges in Real-World and Clinical Settings (MoCha) 2026. arXiv:2608.29928.

Results — gains hold across almost the whole cohort

Per-infant change, base → fine-tuned (ours). Each line is one infant; y inverted where lower is better, so up = better in every panel.



Per-infant change on the 11 held-out infants, base → fine-tuned (*up is better*): **PCK improves on all 11** and every improvement is significant (paired Wilcoxon, $p \leq 0.005$).

Quantitative — 11 held-out infants

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)**

Multi-view consistency rows: means; *teacher agreement* rows: medians (σ_{NIQR}). Sapiens is the *teacher*, so it bounds the 2D rows.

- Prior benchmark: reprojection $39.8 \rightarrow 36.7$ px, GC@10 $0.31 \rightarrow 0.36$.

Qualitative

Three held-out infants; labelled joint = largest gain toward the Sapiens target (black star). **Shape artifacts persist** — adult-proportioned face, over-narrow torso, absent from the pseudo-labels.

Takeaways

- Adapting a 3D body model to a data-scarce population needs **no infant pose labels and no markers** — a frozen teacher, unlabeled video, and a differentiable render to connect them.
- Gains hold **across almost the whole cohort** and are significant on every headline metric (paired Wilcoxon, $p \leq 0.005$).
- **Open:** the mesh *shape* stays adult-proportioned — MHR's shape space is adult-mean — and the dense render terms contribute only 2% of the converged objective.
- **Next:** an infant-capable shape model (ANNY), and the biomechanical head below, so infant movement is returned as clinically interpretable joint angles.