ChildPose: Foundation for Children Pose Modeling

Jiakai Chen[1,2,*] **Yifan Shen**[1,2,*] **Boyi Li**[1,2,*] **Chuanmiao Dong**[2] **Xu Cao**[1,2,†]

[1]PediaMed AI [2]University of Illinois Urbana-Champaign
[*]Equal contribution [†]Corresponding author

xucao@pediamed.ai

Abstract

Human pose estimation has advanced rapidly driven by large-scale adult-centric models. However, children remain significantly underrepresented due to data scarcity, distinct body morphology, and unique motion dynamics. We introduce `ChildPose`, a modular video foundation model designed for child-centric pose estimation. `ChildPose` transforms pretrained pose encoders into a streaming pediatric model via memory-augmented temporal attention and age-aware co-training. By maintaining a memory bank of adjacent frame representations, the model achieves robust localization under motion blur, occlusion, and atypical poses, while an auxiliary age-prediction objective enforces representations that capture developmental morphology. We curate a diverse, video-based pediatric dataset spanning various age groups and capture conditions, evaluating out-of-distribution generalizability on entirely unseen subjects. Across multiple 2D baselines, `ChildPose` consistently outperforms both off-the-shelf models and child-data fine-tuned variants, while ensuring stable sequence-level predictions. Our results underscore that effective pediatric pose estimation necessitates child-specific modeling beyond standard adult-centric estimators. Project page: https://pediamed.ai/ChildPose/

1 Introduction

Human pose estimation (HPE) is a fundamental and rapidly evolving problem in computer vision and a key representation for action recognition [3, 18, 8, 5, 17, 62, 70], gait analysis [13, 75], human-computer interaction [66, 27, 44], and clinical movement analysis [68, 60, 21, 39, 12, 58, 55]. Recent architectures, from pioneering multi-person frameworks [8] and high-resolution CNNs such as HRNet [61] to transformer-based models such as ViTPose [73, 74] and Sapiens [32], have achieved strong performance on standard benchmarks including COCO [42] and MPII [4]. However, despite the proliferation of large-scale HPE models, their training and evaluation remain dominated by adult-centric benchmarks, leaving open the question of whether current HPE models provide reliable representations for children [7, 26, 6]. This gap is not merely a demographic limitation: as HPE increasingly serves as a foundation for human-centered visual understanding, reliable pose estimation should extend to children, an essential population for building inclusive and generalizable human-centered vision systems.

Children represent a distinct morpho-kinematic domain rather than a mere linear scaling of adult physiology. They exhibit profound anatomical and biomechanical differences, including rapidly evolving head-to-body ratios, limb proportions, and distinct movement patterns across developmental stages [64, 63, 1, 25, 36, 53]. Furthermore, infants and young children frequently adopt poses that are exceedingly rare in adult-centric benchmarks, such as crawling, rolling, or being held by caregivers. Accurately capturing these unique pediatric kinematics is critical, as they serve as vital biomarkers for the early assessment and diagnosis of neuromotor conditions, including cerebral palsy, through methods like the General Movement Assessment (GMA) [7, 11]. Consequently,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

off-the-shelf, adult-centric pose estimators experience a substantial domain shift when applied to pediatric cohorts. In precision-demanding clinical scenarios, this discrepancy forces models to rely on erroneous anatomical priors, manifesting as jittering trajectories and structural hallucinations that violate pediatric biomechanics [26, 6, 23, 71].

Constructing a foundation model for pediatric pose estimation requires addressing three coupled challenges. First, the model should be trained and evaluated on child-centered data, yet collecting large-scale, diverse, and high-quality pediatric pose sequences is heavily constrained by privacy regulations, and the cost of professional clinical annotations [38, 37, 28, 59]. Second, the model should reason over videos rather than isolated frames, since children's movements are often temporally irregular, self-occluded, and visually ambiguous in single images [49, 14]. Third, the model should generalize to unseen children rather than memorizing identities, scenes, or capture conditions. Existing pediatric pose methods remain suboptimal in real child-centered settings, whether based on supervised learning [19, 23, 79], domain adaptation from synthetic source data to in-the-wild target images [29, 33], or fine-tuning models trained on synthetic adult datasets [67] for infant pose estimation [23, 6]. This motivates the central question: *how can we build a child-centric pose foundation model that transfers across backbones, uses temporal context, and generalizes reliably to children unseen during training?*

To this end, we introduce `ChildPose`, a child-centric video pose foundation framework. We first curate the Children Pose dataset, a large-scale child-centered video collection spanning diverse developmental stages, motion patterns, and capture conditions, using a human-in-the-loop active learning annotation pipeline that iteratively identifies model failure cases and prioritizes them for manual correction. Rather than designing a single fixed pose estimator, `ChildPose` is modular: it can instantiate strong pretrained pose backbones, including Sapiens [32], ViTPose [73], HRNet [61], and HRFormer [76], and adapt them into sequence-aware pediatric pose models. To achieve this, `ChildPose` augments each backbone with a streaming memory bank that conditions the current frame on recent observations, forming a memory-augmented spatial-temporal architecture that improves robustness to transient occlusion, motion blur, and ambiguous child poses. In addition, `ChildPose` incorporates age-group prediction as an auxiliary objective, encouraging the shared representation to encode developmental morphology relevant to pediatric pose estimation.

We evaluate `ChildPose` against both off-the-shelf adult-centric pose estimators and their child-data fine-tuned counterparts. Across multiple backbone families, `ChildPose` improves frame-wise keypoint accuracy over standard fine-tuning while producing stable sequence-level predictions under video evaluation. More importantly, we evaluate under a subject-held-out protocol in which all test children are excluded from training. Under this setting, `ChildPose` achieves stronger performance on unseen children, demonstrating that child-centric data and video foundation modeling improve out-of-distribution robustness rather than merely fitting the training subjects.

Overall, our contributions are:

- We identify child-centric pose estimation as an underrepresented and domain-shifted setting for human pose foundation models, highlighting the limitations of adult-centric HPE models for pediatric pose and motion understanding.
- We propose a human-in-the-loop data construction pipeline for data-scarce pediatric pose modeling, iteratively using model failures to guide manual annotation and expand coverage of difficult child poses, occlusions, and motion patterns.
- We introduce `ChildPose`, a modular video pose foundation model that improves child pose estimation over off-the-shelf adult-centric models and fine-tuned baselines by adapting strong pretrained backbones into sequence-aware pediatric pose models through memory-augmented temporal modeling and age-aware co-training.
- We establish a subject-held-out evaluation protocol to test generalization to unseen children, providing direct evidence of out-of-distribution robustness for our child-centric pose foundation modeling.

2 Related Works

Human Pose Datasets. Existing pose datasets have supported strong progress in 2D and video-based human pose estimation, with standard benchmarks such as COCO [42], MPII [4], and PoseTrack [31].

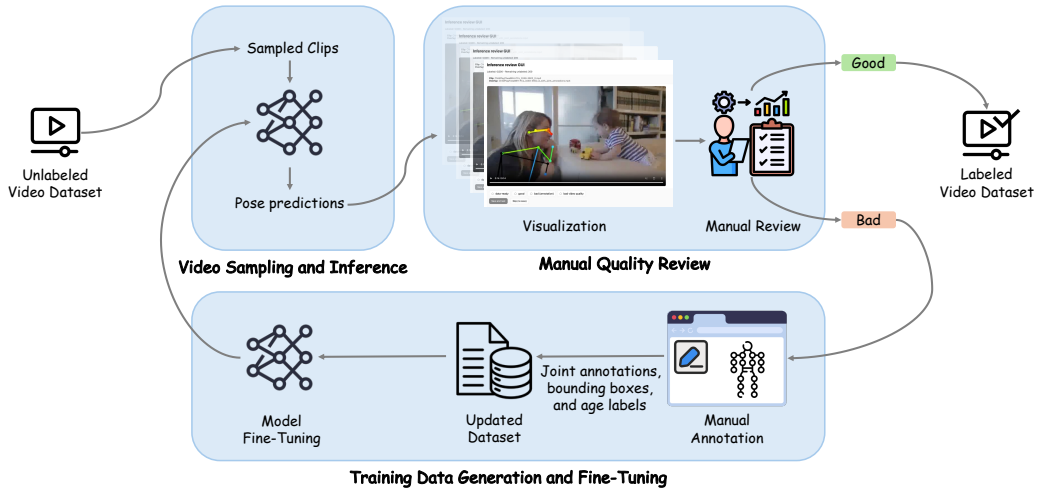


Figure 1: **Human-in-the-Loop Annotation Pipeline.** The current model proposes pose annotations on sampled child videos, and human review separates reliable predictions from failure cases. Reliable samples are added to the labeled set, while hard cases are manually annotated and used to fine-tune the next-round model, progressively shifting the labeled data toward difficult regions of the pediatric pose distribution.

However, these datasets are still mainly adult-centered, which limits their use for evaluating pose models on infants and children. Prior work has noted that children are underrepresented in pose benchmarks and that limited age metadata makes subgroup evaluation difficult [37, 53]. Recent child and infant pose datasets, including MINI-RGBD [20], SyRIP [23], BabyPose [46], InfantNet [26], neonatal pose datasets [15], and child motor development datasets [34], have started to address this gap. These datasets show the need for age-matched data and child-specific evaluation, but coverage across age groups, motion patterns, and recording conditions remains limited.

Human Pose Estimation. Modern pose estimation methods include bottom-up frameworks such as OpenPose [8] and top-down or transformer-based models such as HRNet [61], ViTPose [73], and RTMPose [30]. Recent large-scale pretrained models such as Sapiens [32] further improve transfer for human-centered vision tasks. Beyond keypoint accuracy, recent work also improves the physical plausibility of estimated human poses [41]. However, models trained and evaluated mainly on adult data still show clear limitations when applied to infants and children, where body proportions, poses, and motion patterns differ from adult benchmarks.

Pose Estimation for Children. Pose estimation for children and infants remains more challenging than adult pose estimation because of age-related differences in body proportions, motion patterns, self-occlusion, caregiver interaction, and frequent visual interference from blankets, toys, or clinical equipment [53, 26, 15]. Prior work has shown that models trained mainly on adult datasets often suffer noticeable performance drops in pediatric settings, motivating child-specific adaptations such as transfer learning [23], synthetic infant data generation [20], age aware body models such as SMIL [20], and more inclusive mesh recovery frameworks [10]. Recent datasets including MINI-RGBD [20], SyRIP [23], BabyPose [46], InfantNet [26], BabyView [43], and neonatal or child motor development benchmarks [15, 34] have helped support this direction, but data scarcity and limited diversity are still major bottlenecks. These findings suggest that effective pose estimation for children requires not only stronger backbone models, but also age matched data, child specific evaluation, and modeling choices that better reflect pediatric body structure and real world capture conditions.

3 Video-based Children Pose Dataset

A central obstacle in child-centric pose estimation is the scarcity of video data that captures the diversity of pediatric development, motion, and recording conditions. Existing human pose datasets are largely adult-centered, causing models trained on them to transfer poorly to infants and children [37, 53]. Meanwhile, pediatric pose data is difficult to scale because child videos are sensitive,

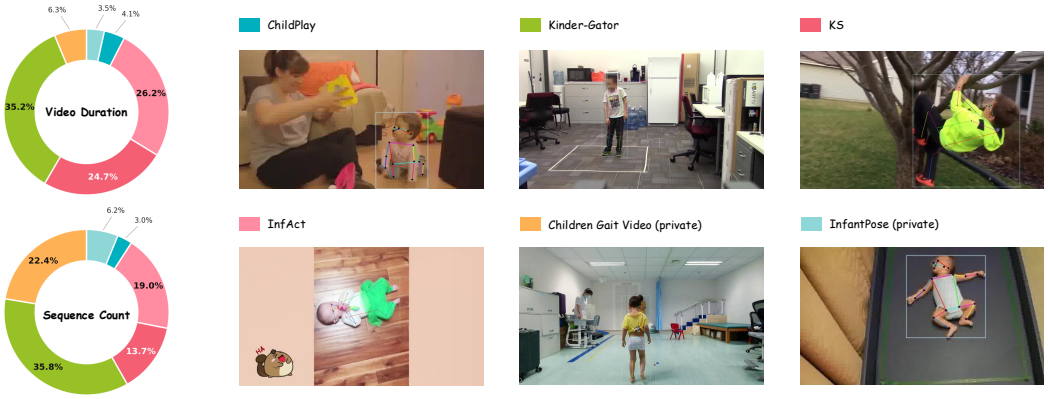


Figure 2: **Dataset Overview.** Source composition of the Children Pose dataset measured by video duration and sequence count, together with representative frames from each source collection. The sources jointly cover diverse child-centered scenes, motion patterns, and recording conditions.

privacy-constrained, and costly to annotate with high-quality keypoint labels [38, 37, 28, 59]. We therefore construct the **Children Pose Dataset** from four public datasets and two private repositories, covering infant postures, child actions, social interaction, and walking-related motion across controlled and in-the-wild settings. To make annotation feasible under this data-scarce setting, we use a human-in-the-loop pipeline that iteratively identifies model failure cases and directs manual effort toward producing reliable annotations for difficult pediatric poses.

Data Sources. Our dataset aggregates diverse public and private sources to ensure comprehensive coverage across developmental stages and environments. We leverage ChildPlay [65] for social interactions and gaze behavior in natural scenes, and InfAct [24] to capture early infant postures and spontaneous movements in home settings. To bridge the gap between infant and adult kinematics, we incorporate Kinder-Gator [2] for structured child motion (ages 5–9) and the KS [47] dataset for unconstrained, in-the-wild pediatric actions. Furthermore, we supplement these with InfantPose [7], which focuses on infant general movement assessment, and the Children Gait Video (CGV) dataset [56, 40] that assesses child gait development from the Edinburgh Visual Gait Score (EVGS) [51]. This multi-source integration provides a robust representation of clinically relevant pediatric motion while maintaining strict privacy standards through rigorous de-identification of all video data. The data collection and annotation process was approved by the Institutional Review Boards (IRBs) of the originating institutions (approval number: 202106202).

Human-in-the-Loop Annotation Pipeline. Our annotation pipeline follows a model-assisted human-in-the-loop design, which is consistent with active learning and model-in-the-loop data construction [54, 78, 35, 50, 26]. In this setting, model predictions reduce annotation cost, while human review and correction control label quality. We treat model predictions as annotation proposals rather than final ground truth. This distinction is important because pediatric pose data contains age-specific body proportions, self-occlusion, caregiver interaction, and motion patterns that are not well covered by adult-centric pose datasets. Since fully manual annotation of large-scale pediatric videos is prohibitively expensive, we use an iterative pipeline that treats model failures as evidence of distributional mismatch. Easy frames add limited new information, while rare child poses, occlusions, and motion patterns require focused human correction.

As shown in Figure 1, we first sample a subset of clips from the unlabeled video corpus. In each round, the current model predicts joint annotations on the sampled clips, and the predictions are rendered as skeleton-overlay videos for manual quality review. A clip is accepted only when the reviewer verifies that the visible keypoints are anatomically plausible, temporally stable, and correctly aligned with the target child across the clip. Samples with visible pose errors, missing visible joints, left-right swaps, unstable keypoints, incorrect child boxes, severe occlusion, or ambiguous body configuration are not directly accepted. For these cases, we sample representative frames as hard cases and send them for manual annotation, where annotators produce joint annotations, object bounding boxes, and age labels. We also filter out videos with extremely low resolution, severe occlusion, or out-of-frame subjects, where reliable pose annotation is not feasible. The manually annotated hard cases are added back to

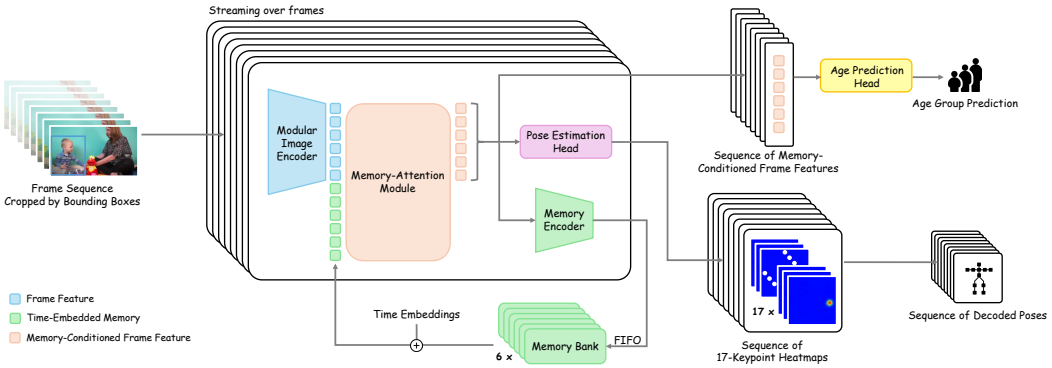


Figure 3: **Architecture of `ChildPose`.** `ChildPose` processes child-centered frame sequences in streaming order. Frame features from a modular image encoder attend to time-embedded memories through the Memory-Attention Module. The memory-conditioned features are decoded into 17-keypoint heatmaps, used for auxiliary age group prediction, and written back into a FIFO memory bank for future frames.

the training set to train the next-round model. Across rounds, this procedure adaptively expands the labeled set toward difficult regions of the pediatric pose distribution and produces reliable annotations.

More broadly, this pipeline provides a general paradigm for data construction under distribution shift: use an imperfect source-domain model to expose target-domain failure modes, prioritize those failures for human annotation, and iteratively move the labeled data toward the target distribution while balancing automation and human oversight.

Dataset Summary. Our dataset contains 5,943 video clips from six sources, totaling 25.6 hours of child-centered video. We sample videos at 5 frames per second and construct 8-frame sequences to match the sequence input used by `ChildPose`. For quality control, we filter out sequences whose average keypoint confidence is below 0.7, where confidence is computed as the mean confidence over all keypoints and all frames in the sequence. We also use SAM3 [9] to obtain child bounding boxes and match them with pose annotations, enabling top-down pose estimation models to use consistent person detections. After filtering, the dataset contains 1,176,332 quality-controlled 8-frame sequences with COCO-style 17-keypoint pose annotations [42] and child instance bounding boxes, half of which also include age labels. We provide visualizations of the data distribution and representative examples in Figure 2. We split the data into training, validation, and test sets with an 8:1:1 ratio. To ensure that evaluation measures generalization to new subjects rather than memorization of training identities, we construct a subject-held-out test split in which every child identity in the test set is absent from the training set. Further details of the dataset are shown in Appendix A.

4 Method

`ChildPose` is a modular video pose foundation model framework for child-centric pose modeling. It follows three principles. First, it is *backbone-modular*: the image encoder can be instantiated with different pretrained pose models and converted into a common token representation, allowing the same child-centric video modeling framework to be applied across backbone architectures. Second, it is *sequence-aware*: frames are processed in temporal order, and a first-in-first-out (FIFO) memory conditions each frame on recent observations to reduce ambiguity from occlusion, motion blur, and single-frame prediction instability. Third, it is *age-aware*: an auxiliary sequence-level age prediction task encourages the shared representation to encode developmental information relevant to pediatric pose estimation. Given a child-centered sequence, `ChildPose` outputs pose heatmaps for all frames and is trained through a staged curriculum that first stabilizes frame-wise pose prediction, then introduces video regularization and age-aware co-training.

4.1 Problem Formulation

We formulate child pose estimation at the frame-sequence level. From each child-centered video clip, we extract 8 consecutive frames, denoted as $\{I_t\}_{t=1}^{T}$ with $T = 8$. Each frame is cropped around

the target child using the associated bounding box with a fixed margin and resized according to the input resolution required by the chosen encoder, yielding $I_t \in \mathbb{R}^{3\times H\times W}$. Our goal is to estimate COCO-style 2D pose annotations with $K = 17$ keypoints [42] for every frame in the sequence. `ChildPose` processes frames in temporal order and outputs a heatmap tensor $\hat{H}_t \in \mathbb{R}^{K\times H_o\times W_o}$ for each frame, from which joint coordinates $\hat{p}_{t,k} \in \mathbb{R}^2$ and confidence scores $c_{t,k}$ are decoded. For sequences with available age labels, we also predict a sequence-level age group as an auxiliary co-training task. The age prediction is used to improve the shared representation and is not required for pose inference.

4.2 Model Architecture

Modular Image Encoder. For each frame I_t, `ChildPose` first extracts a dense frame feature $X_t = E_\theta(I_t)$ using an image encoder E_θ. The encoder is modular and can be instantiated with different pretrained pose backbones, including Sapiens [32], ViTPose [73], HRNet [61], and HRFormer [76]. For each pretrained pose model, we remove the task-specific heatmap prediction layers when present and use the intermediate spatial representation as the encoder output. These backbones provide strong frame-level spatial and pose-structural representations, which serve as the foundation for subsequent temporal conditioning. This modular design allows the same frame-sequence modeling components to be paired with different backbone architectures, so child-centric pose modeling is not tied to a single encoder family. The extracted feature $X_t \in \mathbb{R}^{H'\times W'\times D}$ is then passed to the Memory-Attention Module.

Memory Encoder and Memory Bank. To provide temporal context, `ChildPose` maintains a temporal memory bank for each sequence. After processing frame τ, the memory-conditioned representation $\widehat{X}_\tau$ is projected by a memory encoder into a memory state $M_\tau \in \mathbb{R}^{H_m\times W_m\times D_m}$, which is stored for future frames. The memory bank is maintained as a first-in-first-out (FIFO) queue: at time t, the model retrieves the most recent memories $\{M_\tau\}_{\tau<t}$, and the oldest memory is discarded once the bank exceeds its capacity. We add learnable relative time embeddings $e_{t-\tau}$ to the retrieved memories to obtain time-aware memories $\widehat{M}_\tau = M_\tau + e_{t-\tau}$. In practice, we keep the most recent $N = 6$ memories, which provides short-term temporal context while keeping computation bounded.

Memory-Attention Module. The Memory-Attention Module performs temporal conditioning by allowing the current frame representation to attend to previously stored memories. Given the current frame feature X_t and the time-aware memory set $\widehat{\mathcal{M}}_t = \{\widehat{M}_\tau\}_{\tau<t}$, we apply a memory-attention block containing three operations: self-attention over X_t to capture dependencies among current-frame tokens, cross-attention in which queries are projected from X_t while keys and values are projected from $\widehat{\mathcal{M}}_t$, and a feed-forward network. The final output is the memory-conditioned frame representation $\widehat{X}_t \in \mathbb{R}^{H'\times W'\times D}$. It is passed to the pose head and memory encoder, while the sequence of memory-conditioned representations is further used for age prediction.

Pose Estimation Head. The pose estimation head decodes the memory-conditioned frame representation $\widehat{X}_t$ into frame-wise keypoint heatmaps. We use a heatmap-based regression head with convolutional layers, which outputs $K = 17$ spatial heatmaps $\hat{H}_t \in \mathbb{R}^{K\times H_o\times W_o}$ for each frame. Each heatmap corresponds to one keypoint. During inference, the coordinates of joint k are obtained by taking the peak location of its predicted heatmap, and the confidence score is computed as the corresponding peak response:

$$\hat{p}_{t,k} = \arg\max_{(x,y)} \hat{H}_{t,k}(x, y), \qquad c_{t,k} = \max_{(x,y)} \hat{H}_{t,k}(x, y). \tag{1}$$

Age Prediction Head. The age prediction head provides auxiliary sequence-level supervision for age-aware co-training. Given the memory-conditioned representations from all frames, $\{\widehat{X}_t\}_{t=1}^{T}$, we first apply token pooling to obtain one feature vector per frame. These frame-level features are then passed to a lightweight temporal transformer with a learnable `[CLS]` token. The output `[CLS]` representation is fed into an MLP classifier to predict the age group of the sequence. This auxiliary task encourages the shared representation to capture age-related anatomical and morphological cues that are useful for pediatric pose estimation. The age loss is applied only to sequences with available age labels, and the age head is not required for pose inference.

4.3 Training Objectives

`ChildPose` is optimized with a multi-task objective consisting of frame-wise pose supervision, auxiliary temporal regularization, and age-group classification.

Frame-Wise Pose Loss. For each frame t and keypoint k, we supervise the predicted raw heatmap $\hat{H}_{t,k}$ with a Gaussian target heatmap $H_{t,k}$ using MSE:

$$\mathcal{L}_{\text{pose}} = \frac{\sum_{t=1}^{T} f_t \sum_{k=1}^{K} w_{t,k}\, \text{MSE}(\hat{H}_{t,k}, H_{t,k})}{\sum_{t=1}^{T} \sum_{k=1}^{K} \mathbf{1}[w_{t,k} > 0]}, \tag{2}$$

where f_t is the frame weight and $w_{t,k}$ is a supervision weight that masks missing keypoints and downweights uncertain keypoints.

Auxiliary Temporal Regularization. We obtain differentiable joint coordinates $\tilde{p}_{t,k}$ by applying spatial soft-argmax to each predicted heatmap. Let Ω_{temp} be the set of valid consecutive frame-keypoint pairs, where $(t,k) \in \Omega_{\text{temp}}$ indicates that keypoint k is annotated in both frames $t-1$ and t. We define the scale-normalized displacement as

$$d_{t,k} = \frac{\|\tilde{p}_{t,k} - \tilde{p}_{t-1,k}\|_2}{s_t}, \qquad s_t = \frac{1}{2}\left(\sqrt{A_t} + \sqrt{A_{t-1}}\right), \tag{3}$$

where A_t is the ground-truth bounding-box area in crop-pixel space. The jitter loss is

$$\mathcal{L}_{\text{jit}} = \frac{1}{|\Omega_{\text{temp}}|} \sum_{(t,k)\in\Omega_{\text{temp}}} d_{t,k}. \tag{4}$$

To emphasize reliable motion estimates, we further define the pair confidence $q_{t,k} = c_{t,k}c_{t-1,k}$ and use confidence-weighted jitter:

$$\mathcal{L}_{\text{cwj}} = \frac{\sum_{(t,k)\in\Omega_{\text{temp}}} q_{t,k} d_{t,k}}{\sum_{(t,k)\in\Omega_{\text{temp}}} q_{t,k} + \epsilon}. \tag{5}$$

Finally, we use a soft flicker loss to regularize frame-to-frame confidence changes:

$$\mathcal{L}_{\text{sf}} = \frac{1}{|\Omega_{\text{temp}}|} \sum_{(t,k)\in\Omega_{\text{temp}}} |c_{t,k} - c_{t-1,k}|. \tag{6}$$

Age Group Classification Loss. For sequences with age labels, we apply cross-entropy loss over C age groups:

$$\mathcal{L}_{\text{age}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c, \tag{7}$$

where y is the one-hot ground-truth age label and $\hat{y}$ is the predicted age-group probability.

Training Curriculum and Overall Objective. We employ a staged curriculum: (1) pre-training the pose head and backbone on $\mathcal{L}_{\text{pose}}$; (2) enabling the memory bank and adding temporal terms $\{\mathcal{L}_{\text{jit}}, \mathcal{L}_{\text{cwj}}, \mathcal{L}_{\text{sf}}\}$; and (3) co-training with the auxiliary age head. The total objective is a weighted sum:

$$\mathcal{L} = \lambda_{\text{p}}\mathcal{L}_{\text{pose}} + \lambda_{\text{j}}\mathcal{L}_{\text{jit}} + \lambda_{\text{c}}\mathcal{L}_{\text{cwj}} + \lambda_{\text{s}}\mathcal{L}_{\text{sf}} + \lambda_{\text{a}}\mathcal{L}_{\text{age}}. \tag{8}$$

5 Experiments

5.1 Experimental Setup

Evaluation Protocol. All evaluations are conducted on the subject-held-out test split of the Children Pose dataset. Since every child identity in the test set is absent from the training set, this protocol directly tests whether models generalize to unseen children rather than memorizing training identities.

Baselines. We evaluate representative human pose estimation models, including HRNet [61], HRFormer [76], ViTPose [73], and Sapiens [32]. Since these baselines are image-based pose estimators, they are trained and evaluated at the single-frame level. For each applicable backbone, we

Table 1: **Fine-Tuned Frame-Wise Pose and Video Stability Evaluation.** We evaluate 2D pose estimation models after fine-tuning on the Children Pose training split. All models are evaluated on the subject-held-out test set. Frame-wise pose accuracy is reported using COCO-style AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Video stability is reported using Jitter, Confidence-Weighted Jitter (CWJ), and Flicker Rate. Higher is better for AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Lower is better for Jitter, CWJ, and Flicker Rate. We **bold** the best results and underline the runner-ups.

Model	Visual Encoder	Param	Resolution	Frame-wise Pose				Video Stability		
				AP↑	$AP_{0.50}$ ↑	$AP_{0.75}$ ↑	AR↑	Jitter↓	CWJ↓	Flicker↓
HRNet	HRNet-w32_256x192	28.6M	256x192	0.8261	0.9778	0.8877	0.8548	0.0682	0.0673	**0.0000**
	HRNet-w32_384x288	28.6M	384x288	0.8182	0.9768	0.8856	0.8486	0.0678	0.0669	**0.0000**
	HRNet-w48_256x192	63.7M	256x192	0.8209	0.9764	0.8869	0.8505	0.0679	0.0649	0.0100
	HRNet-w48_384x288	63.7M	384x288	0.8341	0.9778	0.8874	0.8602	0.0671	0.0642	0.0074
HRFormer	HRFormer-small_256x192	7.9M	256x192	0.8305	0.9757	0.9012	0.8624	0.0675	0.0648	0.0097
	HRFormer-small_384x288	7.9M	384x288	0.8344	0.9781	0.8922	0.8633	0.0674	0.0643	0.0103
	HRFormer-base_256x192	43.5M	256x192	0.8471	0.9780	0.9031	0.8743	0.0670	0.0645	0.0072
	HRFormer-base_384x288	43.5M	384x288	0.8403	0.9768	0.9023	0.8694	0.0664	0.0640	0.0077
ViTPose	ViTPose-small_256x192	24.3M	256x192	0.8644	0.9782	0.9174	0.8927	0.0657	**0.0634**	0.0070
	ViTPose-base_256x192	90.0M	256x192	0.8644	0.9783	0.9182	0.8921	0.0658	0.0636	0.0069
	ViTPose-large_256x192	0.3B	256x192	0.8759	0.9891	0.9199	0.8984	0.0653	0.0635	0.0045
	ViTPose-huge_256x192	0.6B	256x192	0.8693	**0.9892**	0.9205	0.8926	**0.0652**	0.0635	0.0048
Sapiens	Sapiens-0.3B-Pose	0.3B	1024x768	0.7693	0.9324	0.8423	0.8049	0.0777	0.0713	0.0193
	Sapiens-0.6B-Pose	0.6B	1024x768	0.7006	0.9176	0.7758	0.7535	0.0815	0.0730	0.0219
	Sapiens-1B-Pose	1B	1024x768	0.7967	0.9406	0.8640	0.8313	0.0761	0.0716	0.0130
	Sapiens-2B-Pose	2B	1024x768	0.8040	0.9325	0.8787	0.8397	0.0750	0.0705	0.0136
`ChildPose`	HRNet-w48_384x288	0.1B	384x288	0.8718	0.9771	0.9300	0.8929	0.0706	0.0654	**0.0000**
	HRFormer-base_384x288	82.4M	384x288	0.8723	0.9777	0.9280	0.8980	0.0690	0.0647	**0.0000**
	ViTPose-base_256x192	0.1B	256x192	**0.8956**	0.9765	**0.9409**	**0.9167**	0.0688	0.0651	**0.0000**
	Sapiens-0.3B-pretrain	0.3B	1024x1024	0.8498	0.9754	0.8988	0.8715	0.0725	0.0660	**0.0000**
	Sapiens-0.6B-pretrain	0.6B	1024x1024	0.8503	0.9754	0.9059	0.8754	0.0712	0.0650	**0.0000**
	Sapiens-1B-pretrain	1B	1024x1024	0.8720	0.9754	0.9276	0.8953	0.0688	0.0638	**0.0000**

report two settings: direct inference using off-the-shelf pretrained checkpoints, and inference after standard single-frame fine-tuning on the Children Pose training split. All models are evaluated in a top-down setting using the same child bounding boxes for cropping at inference; for fine-tuning, the same boxes are also used during training. Thus, the reported results focus on pose estimation rather than person detection. We further evaluate more baselines in Appendix B.

Implementation Details. All experiments are conducted on 4 NVIDIA L40S GPUs. For baseline models, we use the official pretrained weights released by their respective authors and their official fine-tuning scripts when applicable. For `ChildPose`, we instantiate the modular encoder with four backbone families: HRNet-$W48_{384\times288}$, HRFormer-$Base_{384\times288}$, ViTPose-$Base_{256\times192}$, and Sapiens-$pretrain_{1024\times1024}$ at three scales (0.3B, 0.6B, and 1B), where the subscript denotes the input resolution. Each `ChildPose` variant is trained using the staged curriculum described in Section 4 and evaluated on the same subject-held-out test split. We use 8-frame input sequences sampled at 5 FPS for `ChildPose`. Full training and evaluation details are provided in the appendix C.

5.2 Frame-Wise Pose Evaluation

Metrics. We evaluate frame-wise pose accuracy using the official COCO keypoint evaluation protocol [42]. Specifically, we use the COCO API to compute Object Keypoint Similarity (OKS)-based Average Precision and Average Recall over the 17 COCO keypoints. We report AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Additional details of the COCO evaluation protocol are provided in the appendix.

Results. Table 1 reports frame-wise pose estimation results after fine-tuning on the Children Pose training split. `ChildPose` further improves over standard fine-tuning performance across HRNet, HRFormer, ViTPose, and Sapiens encoders. The strongest variant, `ChildPose` with ViTPose-base_256x192 as the visual encoder, achieves the best AP, $AP_{0.75}$, and AR. This shows that memory-augmented frame-sequence modeling and age-aware co-training improve pose estimation beyond backbone fine-tuning alone. Since all results are evaluated on identities excluded from training, the consistent gains across encoder families support that `ChildPose` learns transferable pediatric pose representations.

Table 2: **Ablation Study of `ChildPose` Components.** We evaluate component variants on the subject-held-out test split. "No memory" removes the memory bank. "+Memory" enables memory-based frame-sequence modeling. "+Video loss" adds auxiliary temporal regularization. "+Age" further adds age-aware co-training. We report COCO-style AP, $AP_{0.50}$, $AP_{0.75}$, and AR.

Model	Visual Encoder	Variant	AP	$AP_{0.50}$	$AP_{0.75}$	AR
`ChildPose`	HRNet-w48_384x288	No memory	0.8350	0.9765	0.8890	0.8687
		+Memory	0.8652	0.9770	0.9193	0.8893
		+Video loss	0.8708	0.9765	0.9287	**0.8934**
		+Age	**0.8718**	**0.9771**	**0.9300**	0.8929
	HRFormer-base_384x288	No memory	0.8440	0.9772	0.9044	0.8761
		+Memory	0.8700	0.9775	0.9268	0.8975
		+Video loss	0.8700	0.9775	0.9208	0.8875
		+Age	**0.8723**	**0.9777**	**0.9280**	**0.8980**
	ViTPose-base_256x192	No memory	0.8625	0.9755	0.9188	0.8962
		+Memory	0.8951	0.9757	0.9402	**0.9175**
		+Video loss	0.8950	0.9758	0.9400	**0.9175**
		+Age	**0.8956**	**0.9765**	**0.9409**	0.9167
	Sapiens-0.3B-pretrain	No memory	0.8311	0.9752	0.8867	0.8542
		+Memory	0.8360	0.9751	0.8880	0.8595
		+Video loss	**0.8498**	0.9754	**0.8988**	**0.8715**
		+Age	0.8416	**0.9756**	0.8980	0.8636
	Sapiens-0.6B-pretrain	No memory	0.8422	0.9745	0.8953	0.8674
		+Memory	0.8455	0.9751	0.8957	0.8703
		+Video loss	0.8502	0.9751	0.9055	**0.8761**
		+Age	**0.8503**	**0.9754**	**0.9059**	0.8754
	Sapiens-1B-pretrain	No memory	0.8703	0.9750	0.9268	0.8953
		+Memory	0.8706	0.9753	0.9272	0.8941
		+Video loss	0.8714	**0.9754**	0.9271	**0.8958**
		+Age	**0.8720**	**0.9754**	**0.9276**	0.8953

5.3 Video Stability Evaluation

Metrics. We evaluate video-level behavior using prediction-based temporal diagnostics. Jitter (Equation (4)) and Confidence-Weighted Jitter (CWJ, Equation (5)) are computed from the scale-normalized consecutive-frame displacement defined in Equation (3). These metrics measure inter-frame consistency of predicted trajectories: lower values indicate smoother and more temporally consistent predictions, but do not by themselves imply higher video prediction quality. We therefore interpret Jitter and CWJ as consistency diagnostics rather than direct quality measures.

We additionally report Flicker Rate, which captures practical instability caused by confidence-based filtering. Let $z_{t,k} = \mathbf{1}[c_{t,k} > \tau]$ denote whether keypoint k is retained after thresholding the prediction confidence at frame t, where $\tau = 0.3$ is the confidence threshold. Over the valid consecutive frame-keypoint pairs Ω_{temp}, Flicker Rate is defined as

$$\text{Flicker} = \frac{1}{|\Omega_{\text{temp}}|} \sum_{(t,k)\in\Omega_{\text{temp}}} |z_{t,k} - z_{t-1,k}|. \tag{9}$$

A lower Flicker Rate indicates fewer threshold-induced changes in keypoint availability after confidence filtering and therefore more stable video predictions.

Results. Table 1 also reports video-level stability diagnostics after fine-tuning. `ChildPose` obtains comparable Jitter and CWJ across backbone families while consistently achieving very low Flicker Rate. This indicates that although Jitter and CWJ mainly reflect smoothness and inter-frame consistency, `ChildPose` produces keypoints that remain stable after confidence-based filtering. Together with the frame-wise metrics in Table 1, these results show that `ChildPose` improves pose accuracy while maintaining stable video predictions under subject-held-out evaluation.

We also provide zero-shot inference results using the pretrained pose estimation models in Appendix D, highlighting the gap between child-centered and adult-centric data. The qualitative comparison between zero-shot inference, fine-tuned inference, and `ChildPose` is visualized in Appendix E.

Table 3: **Evaluation on the Public-Only Reproducibility Benchmark.** All models are trained/fine-tuned and evaluated on the subject-held-out split constructed solely from four public sources. Frame-wise pose accuracy is reported using COCO-style AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Video stability is reported using Jitter, Confidence-Weighted Jitter (CWJ), and Flicker Rate. Higher is better for AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Lower is better for Jitter, CWJ, and Flicker Rate. We **bold** the best results and underline the runner-ups.

Model	Frame-wise Pose				Video Stability		
	AP↑	**AP**$_{0.50}$ ↑	**AP**$_{0.75}$ ↑	**AR**↑	**Jitter**↓	**CWJ**↓	**Flicker**↓
ViTPose-B (Zero-shot)	0.6760	0.9692	0.7711	0.7675	0.0724	0.0692	0.0118
ViTPose-B (Fine-tuned)	<u>0.8673</u>	**0.9767**	<u>0.9197</u>	<u>0.9027</u>	**0.0703**	**0.0676**	<u>0.0074</u>
`ChildPose`-ViTPose-B	**0.8761**	<u>0.9726</u>	**0.9338**	**0.9047**	0.0734	<u>0.0689</u>	**0.0000**

5.4 Ablation Study

Module Ablation. We provide ablations in Table 2 for the memory bank, auxiliary temporal regularization, and age-aware co-training. Adding the memory bank consistently improves frame-wise AP over the no-memory variant. The auxiliary temporal regularization and age-aware co-training also provide modest benefits, supporting the contribution of sequence-level modeling and child-aware auxiliary supervision in `ChildPose`.

Reproducibility. We constructed an additional subject-held-out benchmark using only the four public sources. `ChildPose` and its encoder-matched ViTPose-B baselines were trained or fine-tuned and evaluated on the same public-only split. Given the cost of retraining all backbone variants, we report ViTPose-B as a representative public-only experiment. The full paper already evaluates `ChildPose` across four backbone families. This additional experiment is intended specifically to verify reproducibility without relying on private data.

As shown in Table 3, when we restrict both training and evaluation to public data, `ChildPose` improves AP, $AP_{0.75}$, and AR over the fine-tuned ViTPose-B baseline. Its Jitter and CWJ remain comparable, while Flicker is reduced to 0. As discussed in the paper, Jitter and CWJ primarily measure trajectory smoothness and do not necessarily reflect pose accuracy or motion faithfulness, because valid rapid or complex motion can also increase frame-to-frame displacement. We therefore interpret them jointly with COCO pose accuracy and Flicker, which measures threshold-induced instability in keypoint visibility.

Additional Ablation Studies. We also conduct ablation studies on memory capacity (Appendix F) and heatmap decoding and post-processing (Appendix G).

6 Conclusion

We presented `ChildPose`, a modular video pose foundation model for child-centric pose estimation. Motivated by the domain gap between adult-centric pose estimators and pediatric videos, we constructed a video-based Children Pose dataset with a human-in-the-loop annotation pipeline. `ChildPose` adapts pretrained pose backbones into streaming frame-sequence models through memory-augmented temporal conditioning and age-aware co-training. Across multiple backbone families, `ChildPose` improves frame-wise pose accuracy over both off-the-shelf pretrained models and in-domain fine-tuned models, while maintaining stable sequence-level predictions under subject-held-out evaluation. Overall, our results highlight the need to explicitly account for the domain gap between adult and child pose distributions, including age-dependent morphology and child-specific motion patterns, and provide evidence that child-specific video modeling improves generalization to unseen children.

Acknowledgments and Disclosure of Funding

Portions of this work were supported by PediaMed AI and NemX Labs' innovation fund for computational resource.

References

[1] Karen E Adolph, Whitney G Cole, Meghana Komati, Jessie S Garciaguirre, Daryaneh Badaly, Jesse M Lingeman, Gladys LY Chan, and Rachel B Sotsky. How do you learn to walk? thousands of steps and dozens of falls per day. *Psychological science*, 23(11):1387–1394, 2012.

[2] Aishat Aloba, Gianne Flores, Julia Woodward, Alex Shaw, Amanda Castonguay, Isabella Cuba, Yuzhu Dong, Eakta Jain, and Lisa Anthony. Kinder-gator: The uf kinect database of child and adult motion. In *Eurographics (Short Papers)*, pp. 13–16, 2018.

[3] S Mohsen Amiri, Mahsa T Pourazad, Panos Nasiopoulos, and Victor CM Leung. Human action recognition using meta learning for rgb and depth information. In *2014 International conference on computing, networking and communications (ICNC)*, pp. 363–367. IEEE, 2014.

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

[5] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[6] Sarosij Bose, Hannah Dela Cruz, Arindam Dutta, Elena Kokkoni, Konstantinos Karydis, and Amit Kumar Roy Chowdhury. Leveraging synthetic adult datasets for unsupervised infant pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5562–5571, 2025.

[7] Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, and Jianguo Cao. Aggpose: Deep aggregation vision transformer for infant pose estimation. *arXiv preprint arXiv:2205.05277*, 2022.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[9] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[10] Georgios Chatzichristodoulou, Niki Efthymiou, Panagiotis Filntisis, Georgios Pavlakos, and Petros Maragos. Age-inclusive 3d human mesh recovery for action-preserving data anonymization. *arXiv preprint arXiv:2512.05259*, 2025.

[11] Daphné Chopard, Sonia Laguna Cillero, Kieran Chin-Cheong, Annika Dietz, Anna Badura, Sven Wellmann, and Julia E Vogt. Automatic classification of general movements in newborns. In *4th Machine Learning for Health Symposium (ML4H 2024)*. ETH Zürich, 2025.

[12] Hannah L Cornman, Jan Stenum, and Ryan T Roemmich. Video-based quantification of human movement frequency using pose estimation: A pilot study. *PloS One*, 16(12):e0261450, 2021.

[13] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9707–9716, 2023.

[14] Danila Germanese, Sara Colantonio, Marco Del Coco, Pierluigi Carcagnì, and Marco Leo. Computer vision tasks for ambient intelligence in children's health. *Information*, 14(10):548, 2023.

[15] Alex Grafton, Joana M Warnecke, Maxwell Li, Eric He, Lynn Thomson, Kathryn Beardsall, and Joan Lasenby. Neonatal pose estimation in the unaltered clinical environment with fusion of rgb, depth and ir images. *npj Digital Medicine*, 8(1):539, 2025.

[16] Daniel Groos, Lars Adde, Ragnhild Støen, Heri Ramampiaro, and Espen AF Ihlen. Towards human-level performance on automatic pose estimation of infant spontaneous movements. *Computerized Medical Imaging and Graphics*, 95:102012, 2022.

[17] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8):1–34, 2021.

[18] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[19] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[20] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 792–800. Springer, 2018.

[21] Chang Soon Tony Hii, Kok Beng Gan, Nasharuddin Zainal, Norlinah Mohamed Ibrahim, Shahrul Azmin, Siti Hajar Mat Desa, Bart van de Warrenburg, and Huay Woon You. Automated gait analysis based on a marker-free pose estimation model. *Sensors*, 23(14):6489, 2023.

[22] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5699–5708. IEEE, 2020.

[23] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. 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.

[24] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4912–4921, 2023.

[25] Donald F Huelke. An overview of anatomical considerations of infants and children in the adult world of automobile safety design. In *Annual Proceedings/Association for the Advancement of Automotive Medicine*, volume 42, pp. 93, 1998.

[26] Zhipeng Hui, Jiahao Wang, Xun Dong, Guofeng Zhang, Xingrui Wang, Jiawei Peng, Qihao Liu, Xiaoding Yuan, Yi Zhang, Junjie Oscar Yin, et al. Infantnet: A large scale dataset for infant body pose and shape estimation. 2025.

[27] Rongtian Huo, Qing Gao, Jing Qi, and Zhaojie Ju. 3d human pose estimation in video for human-computer/robot interaction. In *International conference on intelligent robotics and applications*, pp. 176–187. Springer, 2023.

[28] Lennart Jahn, Sarah Flügge, Dajie Zhang, Luise Poustka, Sven Bölte, Florentin Wörgötter, Peter B Marschik, and Tomas Kulvicius. Comparison of marker-less 2d image-based methods for infant pose estimation. *Scientific Reports*, 15(1):12148, 2025.

[29] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6780–6789, 2021.

[30] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

[31] Sheng Jin, Xujie Ma, Zhipeng Han, Yue Wu, Wei Yang, Wentao Liu, Chen Qian, and Wanli Ouyang. Towards multi-person pose tracking: Bottom-up and top-down methods. In *ICCV posetrack workshop*, volume 2, pp. 7, 2017.

[32] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pp. 206–228. Springer, 2024.

[33] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *European Conference on Computer Vision*, pp. 603–620. Springer, 2022.

[34] Hye Hyeon Kim, Jin Yong Kim, Bong Kyung Jang, Joo Hyun Lee, Jong Hyun Kim, Dong Hoon Lee, Hee Min Yang, Young Jo Choi, Myung Jun Sung, Tae Jun Kang, et al. Multiview child motor development dataset for ai-driven assessment of child development. *GigaScience*, 12: giad039, 2023.

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

[36] Jürgen Konczak and Johannes Dichgans. The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental brain research*, 117(2):346–354, 1997.

[37] Julienne LaChance, William Thong, Shruti Nagpal, and Alice Xiang. A case study in fairness evaluation: Current limitations and challenges for human pose estimation. In *Association for the Advancement of Artificial Intelligence 2023 Workshop on Representation Learning for Responsible Humancentric AI (R2HCAI), Washington, DC*, volume 1, 2023.

[38] Curtis P Langlotz, Bibb Allen, Bradley J Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S Cook, Adam E Flanders, Matthew P Lungren, David S Mendelson, Jeffrey D Rudie, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 291(3):781–791, 2019.

[39] Ameur Latreche, Ridha Kelaiaia, Ahmed Chemori, and Adlen Kerboua. Reliability and validity analysis of mediapipe-based measurement system for some human rehabilitation motions. *Measurement*, 214:112826, 2023.

[40] Boyi Li, Yifan Shen, Houze Yang, Xu Cao, Guojun Yun, Li Gao, Turong Chen, Long Xu, Jianguo Cao, and Meihuan Huang. The 1st ai children challenge. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2026.

[41] Zhengyuan Li, Zeyun Deng, Yifan Shen, Liangyan Gui, Miaolan Xie, Joseph Campbell, Xifeng Gao, Kui Wu, Zherong Pan, and Aniket Bera. Poseshield: Neural collision fields for human self-collision resolution. In *European Conference on Computer Vision*, pp. 410–428. Springer, 2026.

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[43] Bria Long, Robert Z Sparks, Violet Xiang, Stefan Stojanov, Zi Yin, Grace E Keene, Alvin WM Tan, Steven Y Feng, Chengxu Zhuang, Virginia A Marchman, et al. The babyview dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *arXiv preprint arXiv:2406.10447*, 2024.

[44] Enrico Martini, Harshil Parekh, Shaoting Peng, Nicola Bombieri, and Nadia Figueroa. A robust filter for marker-less multi-person tracking in human-robot interaction scenarios. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 424–429. IEEE, 2024.

[45] Iurii Medvedev, Farhad Shadmand, and Nuno Gonçalves. Young labeled faces in the wild (ylfw): a dataset for children faces recognition. In *2024 IEEE 18th international conference on automatic face and gesture recognition (FG)*, pp. 1–10. IEEE, 2024.

[46] Lucia Migliorelli, Sara Moccia, Rocco Pietrini, Virgilio Paolo Carnielli, and Emanuele Frontoni. The babypose dataset. *Data in brief*, 33:106329, 2020.

[47] Sanka Mohottala, Sandun Abeygunawardana, Pradeepa Samarasinghe, Dharshana Kasthurirathna, and Charith Abhayaratne. 2d pose estimation based child action recognition. In *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*, pp. 1–7. IEEE, 2022.

[48] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*, pp. 137–144, 2016.

[49] Mikkel Damgaard Olsen, Anna Herskind, Jens Bo Nielsen, and Rasmus R Paulsen. Body part tracking of infants. In *2014 22nd International Conference on Pattern Recognition*, pp. 2167–2172. IEEE, 2014.

[50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[51] Heather S Read, M Elizabeth Hazlewood, Susan J Hillman, Robin J Prescott, and James E Robb. Edinburgh visual gait score for use in cerebral palsy. *Journal of pediatric orthopaedics*, 23(3): 296–301, 2003.

[52] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3414–3421, 2013.

[53] Giuseppa Sciortino, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo, and Cosimo Distante. On the estimation of children's poses. In *International conference on image analysis and processing*, pp. 410–421. Springer, 2017.

[54] Burr Settles. Active learning literature survey. 2009.

[55] Yifan Shen, Chuanmiao Dong, Zhuoqing Zhong, Pei Tian, Tianjiao Yu, Jiateng Liu, Bowen Fang, Xinzhuo Li, Yuanzhe Liu, Zhengyuan Li, et al. Position: Multimodal llms should learn from children. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5619–5630, 2026.

[56] Yifan Shen, Boyi Li, Meihuan Huang, Yuanzhe Liu, Xu Cao, Jinyang Jin, Zhengyuan Li, Anglin Liu, Junho Kim, Jingyuan Zhu, et al. Decoding children's gait behavior. In *European Conference on Computer Vision*, pp. 113–136. Springer, 2026.

[57] Yifan Shen, Jiawen Zhang, Jian Xu, Junho Kim, Ismini Lourentzou, Xu Cao, and Meihuan Huang. Evaluating cognitive age alignment in interactive ai agents. *arXiv preprint arXiv:2605.17894*, 2026.

[58] Yifan Shen, Qinghao Zhang, Yuner Zhang, Jiaming Zhou, Houze Yang, Yuhan Wang, Jingyuan Zhu, Onkar Kishor Susladkar, Yu Tian, Meihuan Huang, et al. A survey on embodied ai for pediatrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5605–5618, 2026.

[59] Ameur Soualmi, Christophe Ducottet, Hugues Patural, Antoine Giraud, and Olivier Alata. A 3d pose estimation framework for preterm infants hospitalized in the neonatal unit. *Multimedia Tools and Applications*, 83(8):24383–24400, 2024.

[60] Jan Stenum, Cristina Rossi, and Ryan T Roemmich. Two-dimensional video-based analysis of human gait using pose estimation. *PLoS computational biology*, 17(4):e1008935, 2021.

[61] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

[62] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

[63] David Sutherland, Richard Olshen, and Edmund Biden. *The development of mature walking*. Cambridge University Press, 1988.

[64] David H Sutherland, RICHARD Olshen, LES Cooper, and S LY Woo. The development of mature gait. *Jbjs*, 62(3):336–353, 1980.

[65] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20935–20946, 2023.

[66] Tsung-Han Tsai, Chih-Chi Huang, and Kung-Long Zhang. Design of hand gesture recognition system for human-computer interaction. *Multimedia tools and applications*, 79(9):5989–6007, 2020.

[67] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 109–117, 2017.

[68] Aditya Viswakumar, Venkateswaran Rajagopalan, Tathagata Ray, and Chandu Parimi. Human gait analysis using openpose. In *2019 fifth international conference on image information processing (ICIIP)*, pp. 310–314. IEEE, 2019.

[69] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364, 2020.

[70] Tongyan Wang, Zhengyuan Li, Muhan Lin, Shengyang Luo, Yifan Shen, Aniket Bera, Baijian Yang, and Yingjie Victor Chen. Finemola: towards fine-grained motion-language alignment from clip-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5510–5519, 2026.

[71] Megan A Witherow, Manar D Samad, Norou Diawara, Haim Y Bar, and Khan M Iftekharuddin. Deep adaptation of adult-child facial expressions by fusing landmark features. *IEEE Transactions on Affective Computing*, 15(3):847–858, 2023.

[72] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European conference on computer vision*, pp. 472–487. Springer, 2018.

[73] 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.

[74] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023.

[75] Dingqiang Ye, Chao Fan, Zhanbo Huang, Chengwen Luo, Jianqiang Li, Shiqi Yu, and Xiaoming Liu. Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. *arXiv preprint arXiv:2505.18132*, 2025.

[76] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34:7281–7293, 2021.

[77] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7091–7100. IEEE, 2020.

[78] Chao Zheng, Xu Cao, Kun Tang, Zhipeng Cao, Elena Sizikova, Tong Zhou, Erlong Li, Ao Liu, Shengtao Zou, Xinrui Yan, et al. High-definition map automatic annotation system based on active learning. *AI Magazine*, 44(4):418–430, 2023.

[79] Zhuoran Zhou, Zhongyu Jiang, Wenhao Chai, Cheng-Yen Yang, Lei Li, and Jenq-Neng Hwang. Efficient domain adaptation via generative prior for 3d infant pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 41–49, 2024.

[80] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15039–15053. IEEE, 2023.

A Details of the Children Pose Dataset

A.1 Data Sources

ChildPlay. ChildPlay (CC BY-NC 4.0) [65] contains short video clips collected from YouTube and focuses on children's gaze behavior and social interaction in natural scenes. The videos include children interacting with adults and peers in environments such as kindergartens and therapy-related settings. We use the unlabeled video clips from ChildPlay to increase coverage of natural social interaction, occlusion, and background variation.

InfAct. InfAct (Non-Commercial Research License) [24] is designed for infant posture and action recognition in home videos. It captures early infant behaviors such as sitting, crawling, rolling, and spontaneous movement in daily environments. We use the unlabeled video frames from InfAct to improve coverage of early developmental stages and infant-specific poses.

Kinder-Gator. Kinder-Gator (Unspecified License) [2] contains motion data from children and adults performing natural actions. The child subjects are between 5 and 9 years old, providing structured examples of young-child movement. We use Kinder-Gator to help bridge the gap between infant motion and adult motion in our corpus.

KS. KS (Unspecified License) is a child-centered subset derived from action videos where children are the main performers [47]. The videos come from unconstrained online sources and include diverse action categories, backgrounds, viewpoints, and camera conditions. We use KS to increase coverage of in-the-wild child actions.

InfantPose. InfantPose (Private) [7] contains a large-scale collection of infant movement frames recorded with General Movements Assessment (GMA) devices, focusing on newborn supine position poses. The dataset includes 20,748 images annotated by professional clinicians with 21 refined keypoints, such as fingers, toes, and navels, as well as 15 million unlabeled frames from 5,187 videos. We use InfantPose to increase the data for fine-grained movement detection.

Children Gait Video (CGV). CGV (CC-BY-NC 4.0) [56, 40] contains 339,236 frames from over 1,100 videos of 110 subjects aged 3 to 17, which captures children's gait behaviors from multiple viewpoints and focuses on the fine-grained analysis of pediatric motor patterns. The dataset includes rich frame-level annotations such as 2D anatomical keypoints, bounding boxes, and instance segmentation masks, alongside expert clinical assessments using the Edinburgh Visual Gait Score (EVGS) [51] to aid in the diagnosis of developmental disorders like cerebral palsy. We use the CGV dataset to enlarge the data on subtle and phase-dependent kinematic deviations.

A.2 Age Groups

Overall, the Children Pose dataset has six groups: infant (<1), toddler (1–2), pre-school (3–5), child (6–9), teenager (10–17), and adult ($\geq$ 18). Reliable labels cover 487,397 of 976,551 training windows (49.91%) and 759 of 3,061 training clips (24.80%). The latter includes 30 mixed clips containing more than one age state across tracked instances or sequences. We exclude samples without reliable age metadata only from the age loss, but they still contribute fully to pose training. Within the explicitly labeled windows, the distribution is 18.5% infant, 5.3% toddler, 12.4% pre-school, 55.2% child, 0.1% teenager, and 8.6% adult. These percentages characterize only the labeled subset, not the full dataset; samples without age metadata, including teenage samples, remain in the pose-training corpus.

A.3 Annotation Model and Dataset Composition

Annotation Model. We initialized the annotation pipeline with the 2B-parameter Sapiens pose model using the official checkpoint `sapiens_2b_coco_best_coco_AP_822.pth`. In each round, the current model first generated pose annotations, which were then reviewed by human annotators and divided into accepted and rejected cases. Accepted predictions were directly added to the labeled dataset. Rejected predictions typically corresponded to difficult cases involving occlusion, motion blur, unusual viewpoints, or atypical poses; these cases were manually re-annotated and added to a cumulative human-labeled dataset. The model was then fine-tuned on this accumulated set and used to initialize the next annotation round. Thus, the annotator model was progressively adapted to challenging child poses throughout the iterative process.

Easy and Challenging Cases. In each round, we randomly sampled 200 video clips from the unlabeled pool, reducing the risk that the annotation process would become concentrated only on difficult examples. We conducted 12 rounds of model prediction, human review, correction, and fine-tuning. By the final round, approximately 95% of sampled clips could be directly accepted, indicating that the annotation model had reached a satisfactory balance between labeling accuracy and human effort; we therefore stopped the iterative annotation and fine-tuning process. The resulting dataset retains representative easy cases while progressively enriching difficult cases involving occlusion, motion blur, unstable or missing keypoints, left–right swaps, unusual viewpoints, and ambiguous body configurations.

The final fine-tuned annotation model was then applied to the remaining clips that had not entered the accepted pool. We removed videos with extremely low resolution, severe occlusion, or out-of-frame subjects, and filtered out 8-frame sequences whose mean keypoint confidence across all joints and frames was below 0.7.

A.4 Data Splitting

Throughout the paper, a *video* denotes a raw source recording, which is segmented into overlapping *clips*; each clip yields multiple 8-frame *sequences* (also referred to as windows). The split below is performed at the video level. The Children Pose dataset contains 5,943 video clips from six sources, totaling 25.6 hours of child-centered video. We sample videos at 5 frames per second and construct 8-frame sequences for sequence-level training and evaluation. To control annotation quality, we filter out sequences whose average keypoint confidence is below 0.7, where confidence is computed as the mean confidence over all keypoints and all frames in the sequence. After filtering, the dataset contains 1,176,332 quality-controlled 8-frame sequences with COCO-style 17-keypoint [42] annotations and child instance bounding boxes. Approximately half of the sequences also include age labels.

We split the dataset into training, validation, and test partitions following an 8:1:1 ratio while enforcing subject-level separation. In total, we use 2,088 videos for training and reserve 522 for validation and testing. The validation and test videos are mutually exclusive from the training videos at the child-identity level: every child identity appearing in validation or test is absent from the training set. This subject-held-out design ensures that evaluation measures out-of-distribution generalization to unseen children rather than memorization of training identities, scenes, or repeated video instances.

A.5 Ethical Considerations and Data Privacy

Public Sources. The four public datasets were used under their original terms: ChildPlay (CC BY-NC 4.0); InfAct (Non-Commercial Research License); Kinder-Gator as publicly released by its originating university lab; KS as annotations over a Kinetics-600 subset, whose underlying videos remain subject to their original copyright and platform terms. We do not relicense or redistribute source videos; for web-derived sources we release only annotations and video identifiers. All public-source use was secondary: no recruitment, intervention, identification, or record linkage. Consent and human-subjects oversight for these sources were handled by their original creators under their respective protocols; our use is limited to the research purposes permitted by each license.

Private Sources. InfantPose and raw video sequences of CGV are private. The two private repositories come from previously published retrospective pediatric studies. In addition to public sources, we incorporate these two de-identified private repositories. The private repositories were provided in de-identified form under approved protocols: faces blurred, names, medical-record numbers, birth dates, contact information, identifying overlays, file metadata, and re-identification keys removed before access. We had no access to clinical records or linkage information.

Consistent with established ethical frameworks for pediatric facial, pose, and behavioral datasets [48, 65, 45, 52, 24, 7], the original collection in both was approved by the originating institutions' IRBs (approval number: 202106202) with documented informed consent, in which parents or legal guardians consented to video recording and to the approved secondary research use conducted here. Besides, all videos were rigorously de-identified prior to use, and no personally identifiable information (PII) is retained to ensure strict privacy compliance. Face blurring removes the primary identifier, but body shape and gait are weak biometric signals, so the residual re-identification risk is nonzero. We mitigate it by keeping raw clinical videos on access-controlled servers and never

Table 4: **Evaluation of Additional Baselines.** All models are evaluated on the subject-held-out test set, and all fine-tuned variants use the same Children Pose training split. Frame-wise pose accuracy is reported using COCO-style AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Video stability is reported using Jitter, Confidence-Weighted Jitter (CWJ), and Flicker Rate. Higher is better for AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Lower is better for Jitter, CWJ, and Flicker Rate. We **bold** the best results and underline the runner-ups.

Model	Frame-wise Pose				Video Stability		
	AP↑	$AP_{0.50}$ ↑	$AP_{0.75}$ ↑	AR↑	Jitter↓	CWJ↓	Flicker↓
MotionBERT + ViTPose-B (Frozen)	0.7875	0.9619	0.8626	0.8480	0.0676	0.0676	**0.0000**
MotionBERT + HRNet-W48 (Frozen)	0.7801	0.9587	0.8531	0.8373	0.0691	0.0691	0.0001
MotionBERT + HRFormer-Base (Frozen)	0.7999	0.9618	0.8684	0.8533	0.0694	0.0694	**0.0000**
MotionBERT + ViTPose-B (Joint fine-tuning)	0.8343	0.9720	0.9129	0.8930	**0.0650**	**0.0650**	**0.0000**
MotionBERT + HRNet-W48 (Joint fine-tuning)	0.8104	0.9694	0.8901	0.8739	0.0652	0.0652	**0.0000**
MotionBERT + HRFormer-Base (Joint fine-tuning)	0.8061	0.9706	0.8937	0.8674	0.0652	0.0652	**0.0000**
SHaRPose-B 384×288 (Zero-shot)	0.7416	0.9620	0.8421	0.8042	0.0696	0.0663	0.0124
SHaRPose-B 384×288 (Fine-tuned)	0.8872	0.9790	0.9379	0.9140	0.0662	0.0637	0.0091
SHaRPose-B 256×192 (Fine-tuned)	0.8770	**0.9792**	0.9363	0.9076	0.0670	0.0643	0.0093
`ChildPose`-ViTPose-B	**0.8956**	0.9765	**0.9409**	**0.9167**	0.0688	0.0651	**0.0000**

releasing them, releasing only appearance-free keypoint and bounding-box sequences, and neither performing nor permitting any external-record linkage.

A.6 Licenses and Responsible Use

Released annotations, including pose sequences, bounding boxes, and age labels, will be licensed CC BY-NC 4.0, limited to subsets whose source terms permit redistribution, with DUA-gated access. Raw private clinical videos will not be released under any terms. All `ChildPose` open-sourced model checkpoints follow the strictest upstream backbone's license (Meta's Sapiens license, CC BY-NC 4.0). All releases carry a responsible-use policy prohibiting re-identification, identity linkage, covert surveillance or monitoring of children, and standalone use for clinical diagnosis, treatment, or other high-stakes decisions.

A.7 Comparison with Existing Pediatric Datasets

Table 5 compares Children Pose with representative pediatric pose and motion resources, ordered approximately by the youngest population they target. Children Pose is distinguished not by exceeding every prior dataset on every scale measure, but by jointly providing real RGB video, temporally ordered pose annotations, a consistent COCO-17 schema, and broader developmental and environmental coverage.

Table 5: **Comparison with Existing Pediatric Pose and Motion Datasets.** The comparison is representative rather than exhaustive.

Dataset	Age coverage	Data	Main limitation for our task
BabyPose [46]	Preterm neonates	16 depth videos; 16,000 frames	Depth-only, single-NICU data with 12 limb joints
In-Motion Poses [16]	Young infants	1,424 RGB videos; 20,000 annotated frames	Narrow developmental period, mostly supine GMA, 19-keypoint schema
MINI-RGBD [19]	Infants up to ∼7 months	12 synthetic RGB-D sequences; 12,000 frames	Synthetic RGB-D with limited appearance and motion diversity
SyRIP [23]	Infants	700 real and 1,000 synthetic images	Image-based; no temporal supervision
Multiview Child Motor Dev. [34]	20–71 months	339 children; three-view RGB videos	Structured tasks in a controlled setup
Children Pose (ours)	Infancy to teenage years, plus limited adult references	5,943 RGB clips; 25.6 hours; six sources; COCO-17	—

Table 6: `ChildPose` **Model Configurations.** We report the input crop resolution, encoder token shape, and whether a prefix token is used. The token shape is shown per sequence as $[B, T, N, D]$.

Variant	Input crop	Embedding shape	Prefix token
`ChildPose`-HRNet	384×288	$[B, T, 256, 1024]$	No
`ChildPose`-HRFormer	384×288	$[B, T, 256, 1024]$	No
`ChildPose`-ViTPose	256×192	$[B, T, 192, 768]$	No
`ChildPose`-Sapiens-0.3B	1024×1024	$[B, T, 4097, 1024]$	Yes
`ChildPose`-Sapiens-0.6B	1024×1024	$[B, T, 4097, 1280]$	Yes
`ChildPose`-Sapiens-1B	1024×1024	$[B, T, 4097, 1536]$	Yes

B Evaluation on Additional Baselines

We further evaluate the following baselines on the same subject-held-out test set, and all fine-tuned variants used the same full training split.

MotionBERT. We evaluate MotionBERT [80], which originally takes a sequence of detected 2D joint coordinates as input and learns temporally contextualized motion representations for downstream tasks such as 3D pose estimation. To adapt it to our setting, we first obtain per-frame COCO-17 keypoints from ViTPose-B, HRNet-W48, or HRFormer-Base, feed the resulting 2D skeleton sequence into MotionBERT, and replace its original prediction head with a 2D keypoint refinement head supervised by ground-truth 2D joints. We evaluate both a frozen-detector setting, where only MotionBERT and the refinement head are trained, and a joint fine-tuning setting, where the detector and MotionBERT are optimized end-to-end.

SHaRPose. We also evaluate SHaRPose, which is an image-based 2D pose estimator rather than a temporal model, and we evaluate it frame-wise.

As shown in Table 4, our proposed `ChildPose`-ViTPose-B achieves superior overall frame-wise accuracy and video consistency across the subject-held-out test set, obtaining the highest AP (0.8956), $AP_{0.75}$ (0.9409), and AR (0.9167) while eliminating confidence flicker (0.0000). Although fine-tuned SHaRPose variants demonstrate competitive performance at higher IoU thresholds, reaching an $AP_{0.50}$ of up to 0.9792 and slightly lower jitter, they suffer from persistent flicker issues (around 0.009). On the other hand, incorporating temporal context via MotionBERT effectively stabilizes predictions—yielding near-zero flicker and competitive jitter—yet even joint end-to-end fine-tuning peaks at 0.8343 AP with ViTPose-B, lagging behind pure image-based fine-tuning due to error propagation from 2D coordinate-level refinement. These results confirm that while temporal modeling aids smoothness, `ChildPose`-ViTPose-B provides the best trade-off by establishing new state-of-the-art spatial precision without compromising temporal confidence stability.

C Implementation Details

C.1 Model Details

All `ChildPose` variants follow the same high-level interface. Given cropped frame sequences of shape $[B, T, 3, H, W]$, the modular encoder outputs frame tokens of shape $[B, T, N, D]$. The original task-specific prediction head of a pretrained pose model is removed when present, and heatmaps are predicted only by the `ChildPose` pose head. Table 6 summarizes the encoder configuration and tensor shapes for each instantiated variant.

For `ChildPose`-Sapiens, the encoder is initialized from Sapiens-0.3B-pretrain, Sapiens-0.6B-pretrain, or Sapiens-1B-pretrain. The input crop is 1024×1024, with patch size 16. This gives a 64×64 patch grid and one additional `[CLS]` token, resulting in $N = 4097$ tokens with dimension $D = 1024, 1280$, and 1536, respectively. Since these checkpoints are used as pretrained encoders rather than pose checkpoints, there is no heatmap head to remove.

For `ChildPose`-HRNet, we use HRNet-W48 with 384×288 input crops. The native HRNet heatmap prediction layer is discarded, and the high-resolution feature map is converted into tokens by adaptive pooling to a 16×16 token grid followed by a 1×1 projection. This gives $N = 256$ tokens with dimension $D = 1024$.

Table 7: **Training Curriculum.** We progressively activate model components and losses during training.

Stage	Trainable modules	Memory	Active losses
Phase 1.1	Pose head	Off	Pose heatmap loss
Phase 1.2	Backbone and pose head	Off	Pose heatmap loss
Phase 1.3	Backbone, pose head, and memory pathway	On with warmup	Pose heatmap loss
Phase 2	Backbone, pose head, and memory pathway	On	Pose + temporal losses
Phase 3.1	Age head	On	Age classification loss
Phase 3.2	All trainable branches	On	Pose + temporal + age losses

For `ChildPose`-HRFormer, we use HRFormer-Base with 384×288 input crops. The wrapper takes the backbone feature map and converts it into a 16×16 token grid, giving $N = 256$ tokens with dimension $D = 1024$.

For `ChildPose`-ViTPose, we use ViTPose-Base with 256×192 input crops. The encoder outputs a 16×12 patch-token grid, giving $N = 192$ tokens with dimension $D = 768$.

Memory Bank Configuration. The memory bank receives encoder tokens of shape $[B, T, N, D]$ and outputs memory-conditioned tokens of the same shape. The memory bank stores encoded feature tensors rather than raw images, heatmaps, or keypoints. It is maintained as a FIFO queue with maximum length 6. At each time step, the model first conditions the current frame tokens on past memories, then encodes the resulting memory-conditioned representation and appends it to the memory bank.

The memory encoder maps current-frame tokens from dimension D to memory dimension $D_m = 512$ using LayerNorm, a linear layer, GELU, and a second linear layer. Retrieved memory entries are augmented with learned relative time embeddings and then projected back to the attention dimension D before cross-attention. In the final experiments, the Memory-Attention Module uses one memory-attention block. Each block contains self-attention over current-frame tokens, cross-attention from current-frame tokens to memory tokens, and a feed-forward network.

Pose Head Configuration. The pose head receives memory-conditioned tokens and reshapes them into a spatial token grid. For backbones with a prefix token, such as Sapiens, the prefix token is removed before reshaping. The pose head applies a convolutional stem followed by two transposed-convolution upsampling layers and a final 1×1 convolution to produce $K = 17$ heatmaps. The resulting heatmap resolutions are 64×64 for HRNet and HRFormer, 64×48 for ViTPose, and 256×256 for Sapiens.

Age Branch Configuration. The age branch takes the sequence of memory-conditioned frame tokens as input. For each frame, it first aggregates feature tokens into a single frame-level representation. When the encoder provides a prefix token, this token is used as a query to attend over the patch tokens. Otherwise, patch-feature tokens are mean-pooled. The resulting sequence of frame-level representations is then processed by a temporal transformer with a learnable `[CLS]` token. The final `[CLS]` representation is passed to an MLP age head, which predicts one of six age groups: infant, toddler, pre-school, child, teenager, and adult. This branch is used only as an auxiliary co-training objective.

C.2 Training Details

Data Preprocessing. We train and evaluate `ChildPose` on 8-frame sequences sampled at 5 FPS. Each frame is cropped around the target child using the associated child bounding box. The bounding box is expanded by a factor of 1.1 and resized to the input crop resolution required by the selected encoder. Cropping and resizing use bicubic interpolation, and input images are normalized with ImageNet mean and standard deviation. Keypoint supervision labels are derived from confidence scores using two thresholds, 0.35 and 0.7, producing missing, uncertain, and confident keypoint labels. Missing keypoints are masked out during heatmap supervision, while uncertain keypoints are downweighted during training.

Table 8: **Computational Overhead and Inference Speed Comparison.** Benchmarked on a single NVIDIA L40S GPU using FP16 precision and batch size 1 at 256×192 resolution across 2,048 video frames. End-to-end latency accounts for initial memory population, while steady-state latency and FPS are measured after the $N = 6$ memory bank is filled. Results are averaged over 5 runs ($\pm$ standard deviation) following 64 warm-up iterations. We **bold** the best efficiency numbers.

Model	Param (M)	End-to-End (ms)	Steady-State (ms)	Steady-State FPS↑
Standard fine-tuning	**89.995**	**4.966 ± 0.030**	**4.962 ± 0.029**	**201.52 ± 1.18**
`ChildPose`	108.173	6.086 ± 0.028	6.107 ± 0.029	163.74 ± 0.78

Training Curriculum. Table 7 summarizes the staged training curriculum. We first stabilize frame-wise heatmap prediction without memory, then fine-tune the encoder, enable memory-based sequence modeling, and finally introduce auxiliary temporal regularization and age-aware co-training.

Optimization. Unless otherwise specified, models are optimized with AdamW. The default learning rate is 10^{-4}, with a smaller backbone learning rate of 10^{-5} when full-backbone fine-tuning is enabled, and a cosine scheduler. The default weight decay is 0.01. We select the best checkpoint using validation pose accuracy for pose-related phases and validation age accuracy for the age-only phase.

Loss Weights and Constants. For the final multitask objective, the pose heatmap loss has weight 1.0, jitter loss has weight 0.2, confidence-weighted jitter loss has weight 0.2, soft flicker loss has weight 0.2, and age classification loss has weight 1.0. The temporal losses use differentiable soft-argmax coordinates with temperature 0.3. The training soft flicker loss is computed directly from adjacent-frame confidence differences and does not use a confidence threshold. The hard confidence threshold $\tau = 0.3$ is used only for evaluation Flicker Rate.

Age-Aware Co-Training. Age prediction provides complementary developmental supervision wherever reliable metadata is available. Using this labeled subset, it improves AP by 0.0010, 0.0023, 0.0006, 0.0001, and 0.0006 for HRNet, HRFormer, ViTPose, Sapiens-0.6B, and Sapiens-1B, respectively. The age branch is training-only and adds no inference-time overhead.

Baseline Training. For HRNet, HRFormer, ViTPose, and Sapiens baselines, direct inference uses publicly released pretrained checkpoints. When applicable, fine-tuned baselines are trained with the official fine-tuning procedures from the corresponding methods. The best fine-tuned baseline checkpoint is selected by validation AP. All baseline models are evaluated in the same top-down setting as `ChildPose`: the same child bounding boxes are used for cropping at inference, and the same boxes are used during fine-tuning when training is performed.

C.3 Evaluation Details

Frame-Wise COCO Keypoint Evaluation. We evaluate frame-wise pose accuracy using the official COCO keypoint evaluation. COCO keypoint evaluation is based on Object Keypoint Similarity (OKS), which plays the same role for keypoints as IoU does for object detection. For a predicted pose and a ground-truth pose, OKS is defined as

$$\mathrm{OKS} = \frac{\sum_{k=1}^{K} \exp\left(-\frac{\|\hat{p}_k - p_k\|_2^2}{2s^2\kappa_k^2}\right) \mathbb{I}[v_k > 0]}{\sum_{k=1}^{K} \mathbb{I}[v_k > 0]},$$

where $\hat{p}_k$ and p_k are the predicted and ground-truth coordinates of keypoint k, s is the object scale computed from the ground-truth person area, κ_k is the COCO keypoint-specific falloff constant, and v_k is the ground-truth keypoint visibility label. Only labeled keypoints with $v_k > 0$ contribute to OKS.

Given OKS scores, COCO computes Average Precision (AP) by matching predicted poses to ground-truth poses at different OKS thresholds. The primary AP metric averages precision over thresholds $0.50{:}0.05{:}0.95$. We also report $\mathrm{AP}_{0.50}$ and $\mathrm{AP}_{0.75}$, which compute AP at OKS thresholds 0.50 and 0.75, respectively. Average Recall (AR) is computed analogously by averaging recall over the same OKS threshold range. In all experiments, we use the official COCO API over the 17 COCO keypoints.

Table 9: **Zero-shot inference frame-wise pose and video stability evaluation.** We evaluate publicly released pretrained 2D pose estimation models on the Children Pose benchmark without fine-tuning. All models are evaluated on the subject-held-out test set. Frame-wise pose accuracy is reported using COCO-style AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Video stability is reported using Jitter, Confidence-Weighted Jitter (CWJ), and Flicker Rate. Higher is better for AP, $AP_{0.50}$, $AP_{0.75}$, and AR. Lower is better for Jitter, CWJ, and Flicker Rate. We **bold** the best results and underline the runner-ups.

Model	Visual Encoder	Param	Resolution	Frame-wise Pose				Video Stability		
				AP↑	$AP_{0.50}$ ↑	$AP_{0.75}$ ↑	AR↑	Jitter↓	CWJ↓	Flicker↓
HRNet	HRNet-w32_256x192	28.6M	256x192	0.6995	0.9397	0.7933	0.7604	0.0764	0.0703	0.0220
	HRNet-w32_384x288	28.6M	384x288	0.6871	0.9411	0.7576	0.7482	0.0775	0.0704	0.0254
	HRNet-w48_256x192	63.7M	256x192	0.7057	0.9504	0.7996	0.7727	0.0753	0.0697	0.0200
	HRNet-w48_384x288	63.7M	384x288	0.6953	0.9485	0.7817	0.7657	0.0756	0.0694	0.0190
HRFormer	HRFormer-small_256x192	7.9M	256x192	0.7099	0.9529	0.8153	0.7740	0.0748	0.0696	0.0228
	HRFormer-small_384x288	7.9M	384x288	0.7090	0.9522	0.7915	0.7581	0.0770	0.0692	0.0518
	HRFormer-base_256x192	43.5M	256x192	0.7200	0.9516	0.8103	0.7736	0.0739	0.0690	0.0208
	HRFormer-base_384x288	43.5M	384x288	0.7141	0.9517	0.7993	0.7683	0.0743	0.0691	0.0202
ViTPose	ViTPose-small_256x192	24.3M	256x192	0.7127	0.9532	0.8157	0.7869	0.0714	0.0670	0.0170
	ViTPose-base_256x192	90.0M	256x192	0.7043	0.9492	0.8006	0.7842	0.0701	<u>0.0666</u>	0.0128
	ViTPose-large_256x192	0.3B	256x192	**0.7277**	<u>0.9580</u>	**0.8206**	**0.8071**	<u>0.0674</u>	**0.0653**	<u>0.0072</u>
	ViTPose-huge_256x192	0.6B	256x192	<u>0.7243</u>	**0.9614**	<u>0.8181</u>	<u>0.8066</u>	**0.0671**	**0.0653**	**0.0065**
Sapiens	Sapiens-0.3B-Pose	0.3B	1024x768	0.6379	0.9109	0.6812	0.7038	0.0872	0.0775	0.0277
	Sapiens-0.6B-Pose	0.6B	1024x768	0.6571	0.9026	0.6755	0.7203	0.0852	0.0762	0.0237

Top-Down Evaluation Setting. All models are evaluated in a top-down setting. The same child bounding boxes are provided to all methods for cropping at inference, and the same boxes are used during fine-tuning when training is performed. Predictions are mapped back to the original image coordinate system before COCO evaluation. Therefore, the reported AP and AR focus on keypoint localization rather than person detection.

Video Stability Evaluation. Video stability metrics are computed over valid consecutive frame-keypoint pairs, where a pair (t, k) is valid if keypoint k is annotated in both frames $t-1$ and t. Jitter and Confidence-Weighted Jitter (CWJ) follow the scale-normalized displacement definitions in the main paper. Flicker Rate also follows the definition in the main paper and uses a confidence threshold $\tau = 0.3$ during evaluation. Unlike the training soft flicker loss, which uses adjacent-frame confidence differences without thresholding, the evaluation Flicker Rate is a hard threshold-based diagnostic that measures threshold-induced changes in keypoint availability.

Aggregation. Frame-wise COCO metrics are computed over the full subject-held-out test set. Video stability metrics are computed over valid adjacent frame-keypoint pairs and averaged over the evaluation set. All methods are evaluated on the same subject-held-out split with the same child bounding boxes.

C.4 Computational Overhead

We benchmarked two ViTPose-B models at 256x192 resolution on a single NVIDIA L40S with FP16 and batch size 1: standard pose fine-tuning without temporal memory, and the full `ChildPose`-ViTPose-B model with an N=6 memory bank. Both models processed the same 16 test videos, with 128 temporally ordered frames sampled from each video at a stride of 6, for a total of 2,048 input frames. Inputs were preprocessed and loaded onto the GPU before timing, and the memory bank was reset at the start of each video.

We report two latency settings. End-to-end latency includes all 128 sampled frames of each video, including the first six frames used to fill the memory bank. Steady-state latency excludes these first six frames from timing and measures only frames 7–128, when the ($N = 6$) memory bank is fully populated. Results are averaged over five runs after 64 warm-up forwards.

As shown in Table 8, introducing temporal memory conditioning in `ChildPose` incurs only modest computational overhead while preserving high-throughput, real-time performance. Relative to the standard fine-tuned ViTPose-B baseline, our complete inference model introduces an additional 18.178M parameters (+20.20%) and increases steady-state latency by 1.145 ms per frame (+23.08%, from 4.962 ms to 6.107 ms). Despite this addition, `ChildPose` comfortably delivers real-time inference at 163.74 FPS on a single NVIDIA L40S GPU. Furthermore, the specialized auxiliary components—such as the training-only age branch (+16.843M parameters) and temporal supervision

Table 10: **Ablation Study on the Memory Capacity N.** We report frame-wise accuracy (AP and AR) under two complementary setups: inference-time truncation of the default $N = 6$ model and full retraining with different memory capacities. Higher is better for both metrics. We **bold** the best results and underline the runner-ups within each setting.

Setting	Capacity (N)	AP↑	AR↑
Inference-time truncation	0	0.8918	0.9150
	1	0.8957	0.9167
	2	**0.8958**	**0.9168**
	4	0.8957	0.9167
	6	0.8956	0.9167
Full retraining	0	0.8625	0.8962
	2	0.8949	0.9159
	4	**0.8956**	**0.9170**
	6	**0.8956**	0.9167
	7	0.8948	0.9164

objectives—impose zero overhead at deployment, confirming that the substantial gains in video consistency and spatial precision do not compromise real-world deployment viability.

D Zero-Shot Evaluation of Pretrained Models

Table 9 reports zero-shot inference results using publicly released pretrained pose estimators without fine-tuning on Children Pose. These results provide a reference for the domain gap between adult-centric pretrained pose models and child-centered video data.

E Qualitative Visualization

Qualitative comparison visualization is shown in Figure 4.

F Ablation Study on Memory Capacity

We selected $N = 6$ as a moderate upper bound on short-term temporal context rather than a uniquely optimal value. We conducted two complementary ablations: inference-time truncation of the original $N = 6$ checkpoint without retraining, and full retraining with different memory capacities.

As shown in Table 10, our ablation experiments demonstrate that `ChildPose` leverages short-term temporal context effectively while exhibiting strong robustness across different memory capacities. Specifically, completely removing the memory bank during training incurs a substantial drop in AP from 0.8956 to 0.8625, highlighting that joint temporal co-learning is essential for optimizing the representation of the encoder and pose head. In contrast, merely truncating the memory at inference time ($N = 0$) results in only a marginal decrease to 0.8918 AP, confirming that the benefits of temporal conditioning are largely preserved within the trained weights, with online temporal retrieval offering an extra performance boost. Crucially, across both inference truncation and full retraining regimes, accuracy quickly saturates around $N \geq 1$ or 2 and remains remarkably stable up to $N = 7$. These findings indicate that `ChildPose` does not rely on a brittle, finely tuned memory capacity or an excessively long temporal horizon; rather, the primary performance gain stems from integrating localized, short-term temporal dynamics directly into the learning framework.

G Ablation Study on Heatmap Decoding and Post-Processing

The reported `ChildPose` results use single-pass hard $\arg\max$ followed by linear crop-to-image coordinate mapping. We do not use UDP [22], quarter-pixel correction [72], flip testing [72, 69], or DARK/Taylor refinement [77]. In particular, no gradient- or Hessian-based approximation is applied at inference. The soft-$\arg\max$ used in the temporal training loss is not used for test

Table 11: **Effect of Heatmap Decoding and Post-Processing Refinements on `ChildPose`.** We evaluate the identical `ChildPose`-ViTPose-B checkpoint on the subject-held-out test set under various decoding strategies without retraining or weight modifications. Higher is better for AP. We **bold** the best performance and underline the runner-up.

Decoder / Post-processing	**AP↑**
Plain argmax (Reported setting)	0.8956
Quarter-pixel refinement	0.8984
DARK refinement	0.9006
Flip + argmax	0.8973
Flip + DARK	**0.9032**

coordinates. In contrast, the official fine-tuned ViTPose-B baseline retains its original MMPose inference pipeline, including UDP preprocessing and coordinate inversion, DARK-UDP with Hessian-based Taylor refinement, and flip testing. Thus, the baseline already uses these standard refinement strategies, whereas the reported `ChildPose` result uses the simpler decoder. To measure their effect on `ChildPose`, we re-evaluated the same checkpoint on the same subject-held-out test set without retraining or modifying its weights:

As shown in Table 11, our reported `ChildPose` baseline does not rely on complex post-processing heuristics to achieve superior accuracy, as it adopts a standard single-pass plain $\arg\max$ decoder with basic linear coordinate mapping, omitting UDP, quarter-pixel shifts, flip testing, and Hessian-based approximations. In contrast, the official ViTPose-B baseline inherently integrates standard test-time refinements, including UDP coordinate inversion, Hessian-based DARK Taylor expansion, and flip testing. When progressively equipping our identical, unretrained `ChildPose` checkpoint with these standard post-processing techniques, performance consistently improves: applying DARK refinement alone raises AP by 0.0050 to 0.9006, while coupling DARK with horizontal flip testing pushes accuracy to 0.9032 AP (+0.0076 AP over the reported baseline). These empirical findings verify that the core performance superiority of `ChildPose` is driven by intrinsic architectural and temporal representation gains rather than post-processing artifacts, with existing inference refinement toolkits offering clear, orthogonal improvements.

H Broader Impacts

`ChildPose` may support research on pediatric movement analysis, child-centered human-computer interaction [57], and assistive video understanding by improving pose estimation for underrepresented child populations. While our model improves robustness, erroneous pose predictions could lead to misinterpretations in clinical settings if relied upon without human oversight. `ChildPose` should not be used as a standalone clinical assessment tool. Deployment in medical, educational, or monitoring settings requires additional validation, fairness analysis, privacy safeguards, and human oversight. Beyond age, dimensions such as skin tone, body habitus, clinical condition, and capture setting are unannotated in our data and therefore unvalidated, so deployment requires a deployment-specific fairness evaluation. Pose models for children also carry dual-use risks such as covert monitoring; such uses are prohibited by our responsible-use policy and limited by releasing only appearance-free annotations.

I Limitations

This work has several limitations that suggest directions for future work. First, our current evaluation focuses on top-down 2D pose estimation with provided child bounding boxes. While this setting isolates pose localization, future work should extend the framework toward richer video understanding tasks, including child action understanding, developmental motion analysis, and multi-task pediatric video modeling beyond pose estimation alone. Second, our temporal losses and video metrics are stability diagnostics rather than complete measures of motion understanding. In particular, lower Jitter or CWJ does not necessarily indicate higher video prediction quality when children undergo rapid or complex motion. Future work should develop evaluation protocols that better capture both temporal consistency and motion faithfulness. Finally, because pediatric video data is highly sensitive, the full dataset cannot be publicly released without appropriate privacy protection and access control.

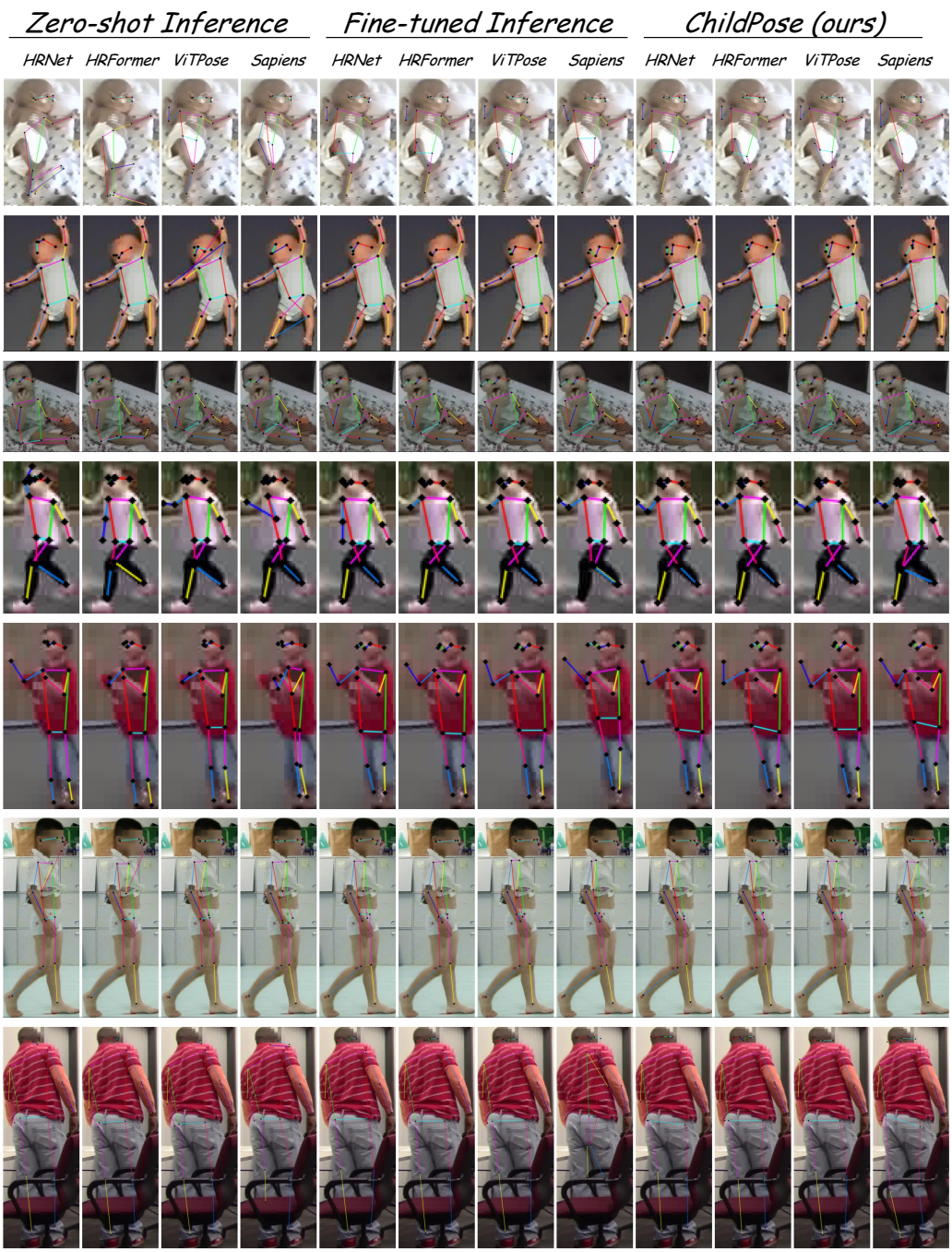


Figure 4: **Qualitative Comparison.** We compare ground-truth annotations with zero-shot inference, fine-tuned inference, and `ChildPose` predictions using different visual encoders. The first column shows the ground truth. The next four columns show zero-shot inference using HRNet, HRFormer, ViTPose, and Sapiens. The middle four columns show fine-tuned inference using the same encoders. The last four columns show `ChildPose` predictions with the corresponding visual encoders.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the proposed ChildPose model, dataset construction, and evaluation protocols, matching the experimental evidence in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Appendix I, including the focus on top-down 2D pose estimation with provided child bounding boxes, the limitations of temporal stability metrics, and the privacy constraints of pediatric data.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper focuses on empirical foundation modeling and dataset curation, and does not contain theoretical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4 describes the architecture and training objectives, and Section 5.1 describes the evaluation protocol, baselines, and implementation. Appendix A details the data sources, filtering, and subject-held-out splits; Appendix C provides model configurations (Appendix C.1), training details including the curriculum, optimizer, and loss weights (Appendix C.2), and evaluation details (Appendix C.3). Appendix 5.4 further reports a public-only benchmark that can be reproduced without private data.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Because pediatric video data is highly sensitive, the full dataset and the raw private clinical videos cannot be publicly released. As described in Appendix A, we will release annotations for the subsets whose source terms permit redistribution under DUA-gated access, together with the ChildPose model checkpoints at https://pediamed.ai/ChildPose/.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data construction, filtering, and splits are described in Section 3 and Appendix A. Training and test details, including the staged curriculum, optimizer, learning rates, weight decay, loss weights, checkpoint selection, and evaluation settings, are given in Section 5.1 and Appendices C.1, C.2, and C.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Pose accuracy and video stability metrics are reported as single values on the large subject-held-out test set, without error bars. Inference latency and throughput (Appendix C) are reported as mean $\pm$ standard deviation over 5 runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5.1 specifies that all experiments are conducted on 4 NVIDIA L40S GPUs, and Appendix C reports parameter counts, inference latency, and throughput on a single NVIDIA L40S GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research carefully handles sensitive pediatric data, emphasizes privacy through de-identification, and notes the need for clinical safeguards.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H discusses positive impacts on pediatric movement analysis and child-centered applications, as well as negative impacts such as misinterpretation of children's movements in clinical settings.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: As described in Appendix A, raw private clinical videos will not be released; released annotations are limited to subsets whose source terms permit redistribution and require DUA-gated access; and all releases carry a responsible-use policy prohibiting re-identification, identity linkage, covert surveillance or monitoring of children, and standalone clinical use.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing datasets and pretrained models are cited in Section 3 and Section 5.1. Appendix A states the license of each dataset (e.g., CC BY-NC 4.0 for ChildPlay and CGV, and the Non-Commercial Research License for InfAct), notes where the license is unspecified (Kinder-Gator and KS), and describes how the original terms are respected. Released checkpoints follow the Sapiens license (CC BY-NC 4.0).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The Children Pose dataset is documented in Section 3 and Appendix A (sources, age groups, annotation process, splits, ethics, and license), and the ChildPose models are documented in Section 4 and Appendix C. Released annotations will be licensed under CC BY-NC 4.0 with DUA-gated access, model checkpoints follow the Sapiens license (https://pediamed.ai/ChildPose/), and raw private videos will not be released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or any direct interaction with participants. The private data are de-identified videos from previously published retrospective pediatric studies, used under approved protocols (Appendix A).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: As described in Section 3, the Ethics statement (Section 6), and Appendix A, the original data collection was approved by the originating institutions' IRBs (approval number: 202106202) with documented informed consent from parents or legal guardians covering video recording and the approved secondary research use. All videos were de-identified before access, and the remaining privacy risks and safeguards are discussed in these sections.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We used LLMs only for text polishing, grammar correction, and code-debugging assistance. They did not contribute to the core methods or ideas.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.