Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Autism classification from public rendered skeleton videos: a participant-level comparative study

Niaz M Doostyar

n.doostyar@kdru.edu.af

Kandahar University

Article

Keywords: Autism spectrum disorder, computer vision, multiple-instance learning, skeleton video, temporal modeling

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11337033/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Public skeleton renderings enable reproducible research on video-based autism classification, but compression and small participant cohorts constrain both learning and evaluation. We investigate ChromaPose-MIL, a compact architecture combining soft chromatic decomposition, factorized spatiotemporal convolutions, temporal attention, and session-level multiple-instance pooling. The classifier is trained from random initialization on an Apple Silicon graphics processor using 160 children, equally divided between autism and typical development. Child-level five-fold cross-validation includes an inner validation partition and a fixed decision threshold. The primary model achieved an area under the receiver operating characteristic curve of 0.605 (conditional 95% interval 0.517–0.692), accuracy of 56.25%, sensitivity of 55.00%, and specificity of 57.50%. A static recording-layout control achieved an area of 0.532. Exploratory nested selection of classifiers on frozen visual features achieved 67.50% accuracy. A full-session neural pooling follow-up achieved 60.63% and did not improve that result. The study demonstrates local computational feasibility and documents how participant-level validation limits conclusions drawn from extensive video data. The observed discrimination and adaptive reuse of one cohort require independent replication before screening use can be considered.

Introduction

Autism spectrum disorder (ASD) is a heterogeneous developmental condition, and timely identification supports access to appropriate services. Surveillance across 16 United States communities estimated that 32.2 per 1,000 eight-year-old children were identified with ASD in 2022; the median earliest known diagnosis was 47 months [1]. These surveillance estimates do not describe the prevalence in a balanced research cohort or in another country. Variation in identification across communities motivates accessible assessment research, while also requiring careful separation of screening assistance from clinical diagnosis.

Video analysis offers a way to quantify movement and interaction patterns that are difficult to summarize consistently by observation alone. In autism research, this possibility has motivated work on gaze, facial behavior, repetitive movement, and social interaction [2]. The scientific task is nevertheless more demanding than separating two sets of recordings. A useful model must generalize to unseen children, and its predictions must be distinguished from the effects of recording conditions, examiner behavior, and the way the cohort was assembled.

Publicly distributed clinical video is often a derived representation rather than the original recording. The OpenPose-ADOS release, derived from Autism Diagnostic Observation Schedule (ADOS) recordings, provides colored skeleton renderings with participant-level labels [3, 4, 5]. These data enable reproducible experiments without requiring access to identifiable face videos, but they remove appearance information and do not provide the original pose coordinates. Consequently, a method designed for raw facial video cannot be transferred to this release without changing the scientific question.

We investigate ChromaPose-MIL, a compact multiple-instance learning (MIL) model that decomposes rendered colors into soft chromatic channels, encodes short motion clips, and aggregates them into a session-level score. The working hypothesis is that an explicit representation of the renderer's colors can reduce the burden on a small network trained from random initialization. This is an engineering hypothesis about the available representation, not a claim that color is a biological marker.

The contributions are a fully specified implementation for rendered skeleton video, a child-level evaluation with matched component removals and simple recording-layout controls, and an audit trail connecting each reported number to predictions and saved split memberships. Established convolution, transformer, and multiple-instance components are credited to their original sources. We distinguish the proposed combination from evidence that it improves discrimination; the latter is determined by the experiments.

Video and behavioral representations

Kojovic et al. studied automated autism prediction from pose-estimation videos using a convolutional and recurrent approach [3]. Natraj et al. extended this direction with complementary video and audio networks [4]. Their public subset is the closest available benchmark for the present work. The release reports 68.75% video-only accuracy on its 80-child test partition [5]. That figure is a published reference point; it is not a paired statistical comparison because the original child-level predictions are unavailable here.

Recent approaches address richer inputs. BASST combines facial appearance with behavioral cues and temporal state-space modeling [6]. Autism-CLIP uses a vision-language framework for caregiver-child interaction videos and includes multicenter evaluation [7]. CAMI-2DNet learns motion representations for assessing imitation [8]. These studies motivate temporal and contextual modeling, but their tasks, modalities, cohorts, and supervision differ. Their reported performance therefore cannot be ranked directly against performance on the rendered ADOS subset.

Pose volumes and session aggregation

OpenPose estimates multiple people's body configurations from images [9]. Coordinate-based methods can then construct joint trajectories or spatial heatmaps; PoseConv3D demonstrates the effectiveness of heatmap volumes for action recognition [10]. The current release instead exposes a compressed colored raster. Our decomposition operates on that raster and retains all visible people. It does not reconstruct missing joints, resolve identity, or distinguish the child from the examiner.

Color-coded skeleton representations precede the present work. Joint Trajectory Maps encode joint trajectories and motion with hue, body-part colormaps, saturation, and brightness [11]. Skeleton Cloud Colorization assigns temporal, spatial, and person-level colors to 3D joints for representation learning [12]. Those methods construct color encodings from coordinates. Here, the input is an already-rendered red-green-blue (RGB) video without released coordinates,

and the decomposition approximately separates its existing palette into soft channels. This difference defines an engineering adaptation; it does not establish a first use of color for skeleton learning or an empirical advantage.

Factorizing spatial and temporal convolutions follows the R(2 + 1)D formulation [13]. Session aggregation uses gated attention multiple-instance learning [14], while temporal token processing follows the transformer family [15]. Neither component is introduced as a new invention. The proposed adaptation concerns their use with soft hue decomposition under session-level diagnostic labels and a constrained local-compute protocol. The component-removal experiments are necessary to establish whether this adaptation has practical value.

Results

ROC-AUC denotes the area under the receiver operating characteristic curve (abbreviated AUC). LR and RF denote logistic regression and random forest; LSTM denotes long short-term memory. Binary cross-entropy is abbreviated BCE.

Primary child-level performance

The complete primary run produced one held-out prediction for each of the 160 children. ROC-AUC was 0.605 (95% interval 0.517–0.692), and accuracy was 56.25% (bootstrap interval 48.13–63.75%). The confusion counts were 44 true positives, 46 true negatives, 34 false positives, and 36 false negatives. The Brier score was 0.252. These figures describe modest discrimination and do not establish a clinically useful operating point.

Table 1
Primary five-fold results, seed zero.

Measure	Value
ROC-AUC	0.605
Accuracy	56.25%
Precision	56.41%
Sensitivity / recall	55.00%
Specificity	57.50%
F1	0.557
Brier score	0.252

Baselines and the supplied split

Table 2
Completed five-fold comparators.

Model	AUC	ACC %
ChromaPose-MIL	0.605	56.25
VGG16-LSTM adaptation	0.632	56.88
Hue dynamics LR	0.600	59.38
Hue dynamics RF	0.647	61.25
Static layout LR	0.532	53.13

The random-forest baseline achieved AUC 0.647 and accuracy 61.25%; the frozen-feature VGG16–LSTM achieved AUC 0.632. Both had higher point estimates of AUC than ChromaPose-MIL. Thus, these experiments do not support a performance advantage for the proposed architecture. The comparisons share the outer participant partitions, but differ in representation and training procedure.

The main model's AUC exceeded the static-layout probe by 0.073. The exploratory unadjusted paired DeLong p-value was 0.164; the exact McNemar p-value was 0.597. This comparison does not establish reliable superiority over the static control. Multiplicity-adjusted comparisons are retained with the machine-readable results.

On the supplied 80-child test partition, the main model achieved AUC 0.574 and accuracy 57.50%. The published release reference is 68.75% accuracy [5]. Differences from that aggregate reference are descriptive because no paired original predictions are available, and the earlier development pilot had already exposed this partition.

Component removals and repeat seeds

Table 3
Completed matched ablations.

Condition	AUC	ACC %
RGB plus motion	0.577	60.63
Fixed hue prototypes	0.604	57.50
Without motion channels	0.557	51.25
Mean session pooling	0.576	56.25
Without temporal transformer	0.614	57.50
Without split-bag terms	0.600	56.25
Without mirror augmentation	0.557	55.63
Spatial mirror only	0.580	53.13
Without auxiliary clip loss	0.593	55.00
Six hue groups	0.610	62.50

Table 4
Split and initialization sensitivity.

Seed	AUC	ACC %
0	0.605	56.25
1	0.593	60.00
2	0.563	55.00

Across the three separately evaluated split-and-initialization seeds, accuracy ranged from 55.00% to 60.00%, with mean 57.08% and sample standard deviation 2.60 percentage points. AUC ranged from 0.563 to 0.605, with mean 0.587 and sample standard deviation 0.021. These summaries describe sensitivity to splitting and initialization; the standard deviations are not confidence intervals. Each seed includes the same children, so these estimates are not pooled as 480 independent observations.

None of the 15 completed seed-zero comparisons with the primary model met the 0.05 threshold after Holm adjustment, for either AUC or paired classification errors. This finding does not establish equivalence. Complete adjusted values and discordant-child counts are included in the reproducibility records.

Training behavior and attention

Training histories are retained for every fold. Lower training objective values alone did not establish stronger held-out discrimination. Supplementary Fig. S2 separates the training objective from inner-validation AUC. Supplementary Fig. S3 shows two examples selected by the first sorted identifier in each diagnostic group, rather than by prediction confidence. Attention weights are relative within a session and do not identify clinically verified behavior.

Exploratory accuracy follow-ups

After the initial results were inspected, we evaluated two additional transfer-learning procedures. Both retain the same five outer child folds and select representations, classifier type, and regularization using four stratified folds within each outer training partition. Each search contains 32 fixed candidates: four representations, logistic regression or a radial-basis support-vector machine (SVM), and C in {0.01, 0.1, 1, 10}. Scaling is fitted within each training partition, and classification uses the fixed zero decision-margin boundary. No outer test labels select a candidate or threshold.

The first search compares VGG16 session means; means, standard deviations, and mean absolute adjacent-frame changes; hue-dynamics descriptors; and concatenated descriptors. The second compares statistics over one, two, and four ordered session partitions; motion differences at lags of 1, 4, 16, and 64 feature frames; the first search's appearance-dynamics representation; and motion/hue fusion. Feature sampling is four frames per second. Both procedures use frozen ImageNet features and therefore do not constitute training the entire classifier pipeline from scratch.

Table 5
Exploratory accuracy follow-ups.

Procedure	ACC %	SENS %	SPEC %
Session statistics	67.50	67.50	67.50
Temporal statistics	67.50	66.25	68.75
Full-session MIL	60.63	57.50	63.75
Hue motion 64 × 48	58.75	58.75	58.75
Hue motion 128 × 96	56.88	60.00	53.75

Both completed procedures achieved 67.50% accuracy, compared with 56.25% for the original network. The session-statistics procedure had a conditional bootstrap accuracy interval of 60.0–75.0%; the temporal procedure had an interval of 60.625–75.0%. Temporal aggregation did not increase accuracy. SVM margins are not calibrated probabilities, and pooled AUC combines fold-dependent score scales; accuracy is the primary endpoint for these searches. Because these follow-ups were designed after earlier results on the same cohort, nested selection does not remove all project-level adaptive reuse. These findings are exploratory and require independent replication.

A third follow-up tested full-session learning without clip-level diagnostic supervision. Contiguous groups of 16 VGG feature frames, or four seconds, are averaged into session tokens. Normalization uses only the optimization children. Each outer fold reserves a stratified 20% inner validation set to select among mean-plus-gated pooling with projection dimension 32 or 64 and mean pooling with dimension 32. All candidates use AdamW, learning rate 0.0005, weight decay 0.03, batch size eight, token dropout 0.2, and bag dropout 0.5. Training uses session BCE for at most 40 epochs with patience ten; candidate and checkpoint selection minimize validation BCE. CPU execution is limited to two threads. All session tokens, including a shorter final token where necessary, are included at training and inference.

Full-session pooling achieved 60.63% accuracy (conditional bootstrap interval 53.13–68.75%), AUC 0.651, sensitivity 57.50%, and specificity 63.75%. Its 46 true positives and 51 true negatives represent 97 correctly classified children. The tested full-session alternative did not improve on the strongest completed feature-statistics procedure. This result concerns the tested configurations and does not establish equivalence between all session-learning approaches.

Two further fixed searches tested spatial motion of rendered hue distributions without pretrained visual features. Sampling at two frames per second, we approximate the 18 fixed hue assignments with 512 angular bins and obtain each color distribution's mass, centroid, spread, and covariance. Spatial summaries exclude observations below 0.2% of total frame chromatic mass. Four representations combine absolute or frame-centered/scale-normalized coordinates with state statistics alone or added differences at lags of 0.5, 2, and 8 seconds. The same 32-candidate grid, four inner folds, outer child folds, and zero-margin boundary are retained. These descriptors combine all visible people and are not recovered anatomical joints or identity-resolved trajectories.

The compact 64 × 48 search achieved 58.75% accuracy and AUC 0.577. Repeating the fixed procedure on the 128 × 96 no-dilation cache achieved 56.88% accuracy and AUC 0.547. Neither improved on the VGG statistical searches. Resolution and dilation changed together, so this is a combined input-representation comparison rather than an isolated resolution ablation. Both failed modifications remain in the research record; all these searches share the limitations of adaptive cohort reuse and fold-dependent margin scales.

Combined higher-resolution neural variant

A fixed exploratory variant retained 128 × 96 RGB frames at eight frames per second using area interpolation without channel-wise dilation. The completed 160-child cache contains 6,080 compressed chunks and occupies 2.745 GiB. A lazy loader retains at most four decoded chunks instead of assembling the entire cohort in memory. Source-fragment fingerprints, cache hashes, frame counts, labels, and occupancy parity were verified before pooled analysis.

The variant uses fixed hue prototypes and concentration, convolutional widths 16, 24, 48, and 80, a 64-dimensional temporal representation, eight clips per bag, four children per batch, learning rate 0.0003, and weight decay 0.03. It retains spatial augmentation but removes reflection, auxiliary clip loss, and split-bag terms, leaving smoothed session BCE. The classifier has 153,076 trainable parameters. The original five child folds, inner validation selection, 40-epoch cap, patience, and fixed 0.5 decision threshold are retained. These simultaneous changes define a combined representation and training comparison, not an isolated test of resolution.

The completed variant achieved 54.38% accuracy (conditional bootstrap interval 46.88–62.50%), AUC 0.585 (conditional interval 0.496–0.673), sensitivity 71.25%, specificity 37.50%, precision 53.27%, and F1 0.610. Confusion counts were TP 57, TN 30, FP 50, and FN 23. It did not improve the primary accuracy or the strongest statistical follow-ups. This outcome limits the tested modification; it does not establish that higher-resolution video is uninformative.

The primary confusion matrix is provided in Supplementary Fig. S1. Complete supplied-split values are provided in Supplementary Table S1.

Complete paired test values and discordant-child counts are provided in Supplementary Tables S2 and S3.

Discussion

The primary result demonstrates that the compact architecture can be trained and evaluated on local Apple Silicon using the complete public cohort. Its observed discrimination is modest. Technical feasibility is therefore better established than clinical usefulness. The result does not justify presenting this model as an accurate diagnostic system or treating its score as a calibrated probability of autism in a clinical population.

The comparison with a static recording-layout model is informative but limited. A stronger temporal model could be exploiting behavior, remaining recording structure, or interactions between both. Conversely, a near-chance static probe would only reject the particular low-dimensional shortcut representation tested here. It would not establish the absence of examiner, camera, or protocol confounding. This distinction is consistent with the broader problem of shortcut learning [16].

The proposed contribution is the adaptation of a differentiable hue decomposition to session-level learning from public rendered skeleton video. Color-coded skeleton representations and the network's constituent mechanisms are established [11], [12]. We make no claim that attention pooling, factorized convolution, label smoothing, consistency regularization, or reflection augmentation is individually novel. The ten matched ablations did not establish a reliable benefit of the tested components after multiplicity adjustment. This result narrows the contribution to a reproducible comparative finding under the specified protocol; it neither demonstrates equivalence nor rules out benefits in larger cohorts or other configurations.

Several explanations could account for limited performance. Downsampling may erase subtle movement; temporal sampling may lose fast events; fixed color assumptions may be inaccurate for the released renderer; and the small number of independent children may constrain learning despite the large number of frames. Auxiliary clip-level supervision also assigns the same diagnostic label to clips that may contain little relevant behavior. These possibilities remain hypotheses, not established causes of the observed errors.

The recordings come from a clinical assessment context and include other visible people. The model may encode the examiner's responses or the structure of the interaction, rather than a child-specific signature. Attention weights describe the model's aggregation and do not provide a causal explanation, an ADOS item score, or a clinician-validated behavioral annotation. Renderer-specific channel perturbations must be interpreted with the same restraint.

The cohort contains ASD and TD groups rather than the full range of developmental presentations encountered in referral practice. No external site, prospective cohort, or independent Afghan population was evaluated. Individual demographic metadata are absent from the local release, preventing subgroup fairness analysis. Scores estimated in a balanced research sample cannot be assumed to retain their predictive values under a different prevalence or referral process.

The proposed classifier is trained from scratch, but its inputs were produced by an existing pose-estimation system. Thus, the complete acquisition-to-score chain is not trained end to end. It cannot directly analyze an unprocessed phone video. Future work should first verify renderer semantics, obtain identity-resolved trajectories where ethically available, and conduct a preregistered evaluation on an independent cohort with clinically relevant comparison groups. A larger model alone would not resolve these validity limitations.

ChromaPose-MIL provides a reproducible implementation for learning participant-level scores from rendered ADOS skeleton videos on local Apple Silicon. The primary five-fold analysis achieved 56.25% accuracy and AUC 0.605; the matched ablations did not establish an advantage for its proposed components. Exploratory nested selection on frozen visual features achieved 67.50% accuracy, with the uncertainty of adaptive reuse of the same cohort. The accompanying predictions, split records, trained models, and recording-layout controls make these conclusions inspectable. Independent clinical cohorts and clinically relevant comparison groups are needed before the resulting research scores can support screening use.

Methods

Data and preprocessing

Cohort and unit of analysis

The analysis uses version 2 of the public OpenPose-ADOS data, DOI 10.5281/zenodo.12658214, distributed under CC BY 4.0 [5]. The local manifest contains 160 child identifiers: 80 labeled ASD and 80 labeled typically developing (TD). The supplied training and testing partitions each contain 40 children per class. All fragments belonging to an identifier remain together throughout training, validation, and testing.

The source study describes children aged 1.1–5.1 years, with 48 males and 32 females in each group [4]. These are published cohort-level descriptors. The local manifest has no individual age, sex, language, cognitive-level, or examiner fields, so demographic adjustment and subgroup performance are not evaluated. ASD labels represent the source investigators' clinical classification; they are not assigned by our preprocessing pipeline.

The local release contains 19,360 MP4 fragments, 121 per child. Ordered fragments are concatenated for decoding. The resulting caches contain approximately 605 seconds per child, with 4,836–4,843 sampled frames. Whole-session array hashes are distinct across all 160 identifiers. This check excludes exact duplicate caches, but cannot independently verify identities or exclude partial duplication across different recordings.

Representation and sampling

FFmpeg samples the concatenated stream at eight frames per second (fps). A channel-wise 3 × 3 dilation is followed by area interpolation to 64 × 48 pixels, preserving thin rendered strokes at a modest computational cost. RGB values are stored as unsigned eight-bit integers and divided by 255 at model input. Dilation, compression, and downsampling can mix neighboring colors; the resulting channels must therefore be interpreted as approximate chromatic densities.

Training clips contain 32 frames, or four seconds. Candidate starts are separated by 16 frames. A clip is retained when at least half its frames have source-frame occupancy above 0.0001; occupancy is the fraction of pixels with a maximum color channel above 40. Sessions with fewer than eight accepted clips

use the unfiltered candidate lattice. The average fraction of effectively empty frames is 1.00%, with substantial session-level variation. The pipeline does not impute poses or infer behavior during occlusion.

Twelve candidate clips are sampled per child per training bag, with replacement only when necessary. Start positions receive up to eight frames of temporal jitter. Spatial scale and translation are constant within a clip, with affine scale in [0.85, 1.00] and normalized translation in [− 0.08, 0.08]. The inherited main configuration applies horizontal reflection with probability 0.5 and a nearest-hue palette permutation. We treat this operation as a renderer-dependent augmentation heuristic; the no-reflection and spatial-reflection-only ablations test its effect. Its anatomical validity is not assumed from the colors alone.

Model and training

Chromatic decomposition

For normalized RGB values r, g, and b, opponent coordinates are u = r − (g + b)/2 and v = √3(g − b)/2. Hue is θ = atan2(v, u), and chroma is ρ = √(u² + v² + 10⁻⁸). Eighteen prototype angles are initialized from the palette implemented in the repository. A soft circular assignment gives each chromatic component p, normalized by the sum over all 18 prototypes. The prototypes and positive concentration κ are learnable in the main model; κ starts at 30.

Pixels with maximum RGB intensity at most 0.02 are assigned zero density. The operation is evaluated on the remaining pixels and scattered into a dense clip tensor. Absolute differences between adjacent density frames form 18 additional motion channels, with a zero initial frame, yielding 36 input channels. These are soft color assignments rather than joint-confidence maps. The fixed-palette, RGB-input, grouped-channel, and no-motion conditions isolate different assumptions of this representation.

$$q_k = exp\left[\kappa \cos(\theta - \mu_k)\right]$$

$$p_k = \rho \frac{q_k}{\sum q_j} \quad (1)$$

Clip encoder and session pooling

The clip encoder starts with a spatial 3 × 3 convolution of stride two and 32 output channels. Three residual blocks increase the widths to 48, 96, and 160. Each block separates a spatial convolution from a temporal convolution, with batch normalization and GELU activation; the final two blocks downsample time by two. Spatial average pooling and a linear projection produce eight 128-dimensional temporal tokens. Eight learned positional embeddings are added before two pre-normalized transformer layers, which use four attention heads, 256-dimensional feed-forward layers, GELU activation, and dropout 0.2. A learned temporal attention pool and layer normalization produce the clip embedding.

Gated multiple-instance pooling with hidden width 64 assigns a nonnegative normalized weight to each clip embedding and forms their weighted sum [14]. A linear classifier produces the session logit. At inference, the model encodes every accepted clip in chunks and then pools the entire session. Thus, inference is invariant to the order of clip embeddings, although the encoder preserves frame order within each clip. It does not model long-range ordering between clips. An auxiliary linear head predicts the child label from individual embeddings during training. The main model contains 578,615 trainable parameters.

Objective and optimization

Let B denote binary cross-entropy (BCE) with logits and yε = 0.95y + 0.025 the smoothed label. Each bag is randomly divided into two equal halves. The total objective combines the full-bag loss, mean auxiliary clip loss, supervision of both half-bags, and a symmetric divergence between their Bernoulli predictions. The executed coefficient on each half-bag supervised term is 0.25. The symmetric divergence D is the average of the two directed Kullback–Leibler divergences, computed with a numerical offset of 10⁻⁶.

Optimization uses AdamW [17], initial learning rate 0.001, weight decay 0.05, gradient-norm clipping at two, and eight children per batch. The schedule multiplies a three-epoch linear warmup factor by a cosine decay over the configured training steps. Three shuffled passes through the training children are sampled per epoch; an incomplete final batch is omitted. Class-alternating ordering keeps batches approximately balanced. Training is capped at 40 epochs.

An inner stratified validation subset contains 20% of the outer training children, rounded upward within each class. For five-fold evaluation this gives 102 optimization children and 26 validation children. Validation is performed every three epochs using every second accepted clip. Early stopping selects the lowest validation cross-entropy, with area under the receiver operating characteristic curve (ROC-AUC, abbreviated AUC) as a tie-breaker, and stops after eight evaluations without improvement. Exponential moving-average weights are evaluated, with decay min(0.99, (1 + s)/(10 + s)) at step s. Encoder computation uses float16 Metal Performance Shaders (MPS) autocast; pooling and losses use float32. Computation runs locally on an Apple M3 Max with 16 CPU cores, a 40-core GPU, and 48 GB of unified memory. Early audited runs recorded Python 3.14.6 and macOS 27; the October continuation records Python 3.14.8 and macOS 27.0.1. Both environments use PyTorch 2.14.0, and exact dependency snapshots and new per-fold environment records are retained.

$$L = B_{bag} + 0.3B_{clip} + 0.25(B_1 + B_2) + 0.5D \quad (2)$$

Experimental design

Validation and comparisons

The primary analysis is stratified five-fold cross-validation at the child level, with 16 children per class in every test fold. Seed zero defines the primary matched comparison; the repeated-seed protocol uses seeds one and two to assess sensitivity to splitting and initialization. Each child contributes exactly one out-of-fold prediction per seed. Classifier selection and early stopping use only the inner training partitions. The threshold is fixed at 0.5 and is not selected on test labels.

The supplied 80/80 split is a secondary exploratory benchmark. A preliminary run on that test set had already been inspected during project development, before the present 40-epoch protocol was fixed. Consequently, it is not an untouched confirmatory holdout. Cross-validation on the same public cohort also cannot substitute for independent external validation. We report this development history because it limits claims about generalization.

Comparators include an adapted frozen ImageNet VGG16 plus long short-term memory (LSTM) classifier, logistic regression (LR) and random forest (RF) on hue-dynamics descriptors, and a static-layout logistic regression probe. Original MP4 fragments are decoded at four frames per second and area-resized directly to 224 × 224 pixels. Frozen VGG16 convolutional features use ImageNet channel normalization, global spatial averaging to 512 dimensions, and float16 cache storage. The adapted recurrent comparator uses these features, 20-frame sequences, a 256-unit LSTM, and mean segment probabilities. It is not an exact reproduction of the published network. The static probe uses a 6 × 8 spatial average of time-averaged maximum-channel intensity plus occupancy mean and standard deviation. Standardization is fitted within each training fold.

The hue-dynamics baselines summarize six fixed channel groups using density and temporal-change statistics and spectral-band energies. They are engineered image descriptors rather than calibrated movement measurements. Logistic regression uses C = 0.1; the random forest uses 500 trees and a minimum leaf size of two. The component-removal protocol retains the main split and training budget while changing the input, palette learning, motion channels, temporal transformer, session pooling, split-bag terms, auxiliary clip supervision, or reflection rule.

Metrics and uncertainty

We report ROC-AUC, accuracy, precision, sensitivity, specificity, F1, and confusion counts. Balanced accuracy equals accuracy in the balanced complete cohort. AUC confidence intervals use DeLong's method [18]; accuracy intervals use 2,000 child-level bootstrap samples. The Brier score summarizes probability error. These intervals are conditional summaries of the available out-of-fold predictions and do not capture all uncertainty from model selection or repeated training [19].

Paired comparisons use two-sided DeLong AUC tests and an exact two-sided binomial version of McNemar's test [18], [20], with a 0.05 decision threshold. Holm adjustment is applied within each protocol and test family across the completed comparisons [21]. Because cross-validation models have overlapping training sets, the paired tests on pooled predictions are exploratory. We therefore interpret effect sizes, intervals, and repeat-seed variation together rather than equating a nominal p-value with a validated clinical improvement.

Ethics

This study is a secondary computational analysis of the publicly available, de-identified OpenPose-ADOS rendered-skeleton dataset. No participants were recruited, no new recordings were collected, and no identifiable source videos or audio recordings were accessed. The original study reports approval by the Ethics Committee of the Faculty of Medicine at the University of Geneva, Switzerland, and written informed consent from a parent or legal guardian for every participating child [4]. No separate ethics committee review, approval or exemption was obtained for the present secondary analysis. No participant re-identification was attempted.

Declarations

Competing interests

The author declares no competing interests.

Funding

This research received no external funding.

Author Contribution

N.M.D. conceived the study, developed the methodology and software, curated the dataset, conducted the experiments and statistical analyses, prepared the figures, wrote and revised the manuscript, and approved the final version.

Acknowledgement

The author thanks the OpenPose-ADOS investigators for collecting and releasing the de-identified rendered-skeleton recordings and diagnostic labels used in this study. The author also acknowledges the developers and maintainers of the open-source scientific software used for data processing, model development and statistical analysis.

Data Availability

The OpenPose-ADOS rendered-skeleton dataset analysed in this study is publicly available on Zenodo at https://doi.org/10.5281/zenodo.12658214. The public code repository is https://github.com/Niaz-M-Doostyar/chromapose-mil-asd. Participant recordings, trained checkpoints, training logs, and participant-level predictions are not deposited in the public code repository.

Code availability

Source code for preprocessing, model development, training, participant-level evaluation, baseline comparisons, ablations and inference, together with tests, dependency specifications and experiment configurations, is publicly available at https://github.com/Niaz-M-Doostyar/chromapose-mil-asd.

References

1. Shaw, K. A. et al. Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years—Autism and Developmental Disabilities Monitoring Network, 16 sites, United States, 2022. *MMWR Surveill Summ.* **74**, 1–22. https://doi.org/10.15585/mmwr.ss7402a1 (2025).
2. de Belen, R. A. J., Bednarz, T. & Sowmya, A. Del Favero, D. Computer vision in autism spectrum disorder research: A systematic review of published studies from 2009 to 2019. *Transl Psychiatry*. **10**, 333. https://doi.org/10.1038/s41398-020-01015-w (2020).
3. Kojovic, N., Natraj, S., Mohanty, S. P., Maillart, T. & Schaer, M. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Sci. Rep.* **11**, 15069. https://doi.org/10.1038/s41598-021-94378-z (2021).
4. Natraj, S., Kojovic, N., Maillart, T. & Schaer, M. Video-audio neural network ensemble for comprehensive screening of autism spectrum disorder in young children. *PLoS ONE*. **19**, e0308388. https://doi.org/10.1371/journal.pone.0308388 (2024).
5. Natraj, S., Kojovic, N., Maillart, T. & Schaer, M. Video-audio neural network ensemble for comprehensive screening of autism spectrum disorder in young children (Openpose ADOS dataset). Zenodo version 2 (2024). https://doi.org/10.5281/zenodo.12658214
6. Bekele, M. W. et al. Multi-cue behavior-aware transformer for autism spectrum disorder screening from facial videos. *Displays* **95**, 103629. https://doi.org/10.1016/j.displa.2026.103629 (2026).
7. Deng, R. et al. Automated screening of autism via a pretrained vision-language model in naturalistic caregiver-child interactions. *Nat. Commun. Adv. online publication*. https://doi.org/10.1038/s41467-026-77721-8 (2026).
8. Kinfu, K. A. et al. Computerized assessment of motor imitation for distinguishing autism in video (CAMI-2DNet). *IEEE Trans. Biomed. Eng.* **73**, 2549–2559. https://doi.org/10.1109/TBME.2025.3637089 (2026).
9. Cao, Z. et al. Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 172–186. https://doi.org/10.1109/TPAMI.2019.2929257 (2021).
10. Duan, H., Zhao, Y., Chen, K., Lin, D. & Dai, B. Revisiting skeleton-based action recognition. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2969–2978 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Duan_Revisiting_Skeleton-Based_Action_Recognition_CVPR_2022_paper.html
11. Wang, P., Li, W., Li, C. & Hou, Y. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowl. -Based Syst.* **158**, 43–53. https://doi.org/10.1016/j.knosys.2018.05.029 (2018).
12. Yang, S., Liu, J., Lu, S., Er, M. H. & Kot, A. C. Skeleton cloud colorization for unsupervised 3D action representation learning. In Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 13423–13433 (2021). https://openaccess.thecvf.com/content/ICCV2021/html/Yang_Skeleton_Cloud_Colorization_for_Unsupervised_3D_Action_Representation_Learning_ICC
13. Tran, D. et al. A closer look at spatiotemporal convolutions for action recognition. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 6450–6459 (2018). https://doi.org/10.1109/CVPR.2018.00675
14. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. In *Proc. Int. Conf. Mach. Learn.* (PMLR) 80, 2127–2136 (2018).
15. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**, 5998–6008 (2017).
16. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673. https://doi.org/10.1038/s42256-020-00257-z (2020).
17. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In Proc. Int. Conf. Learn. Represent. (2019). https://arxiv.org/abs/1711.05101
18. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845. https://doi.org/10.2307/2531595 (1988).
19. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061 (2018).
20. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157. https://doi.org/10.1007/BF02295996 (1947).
21. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

Figures

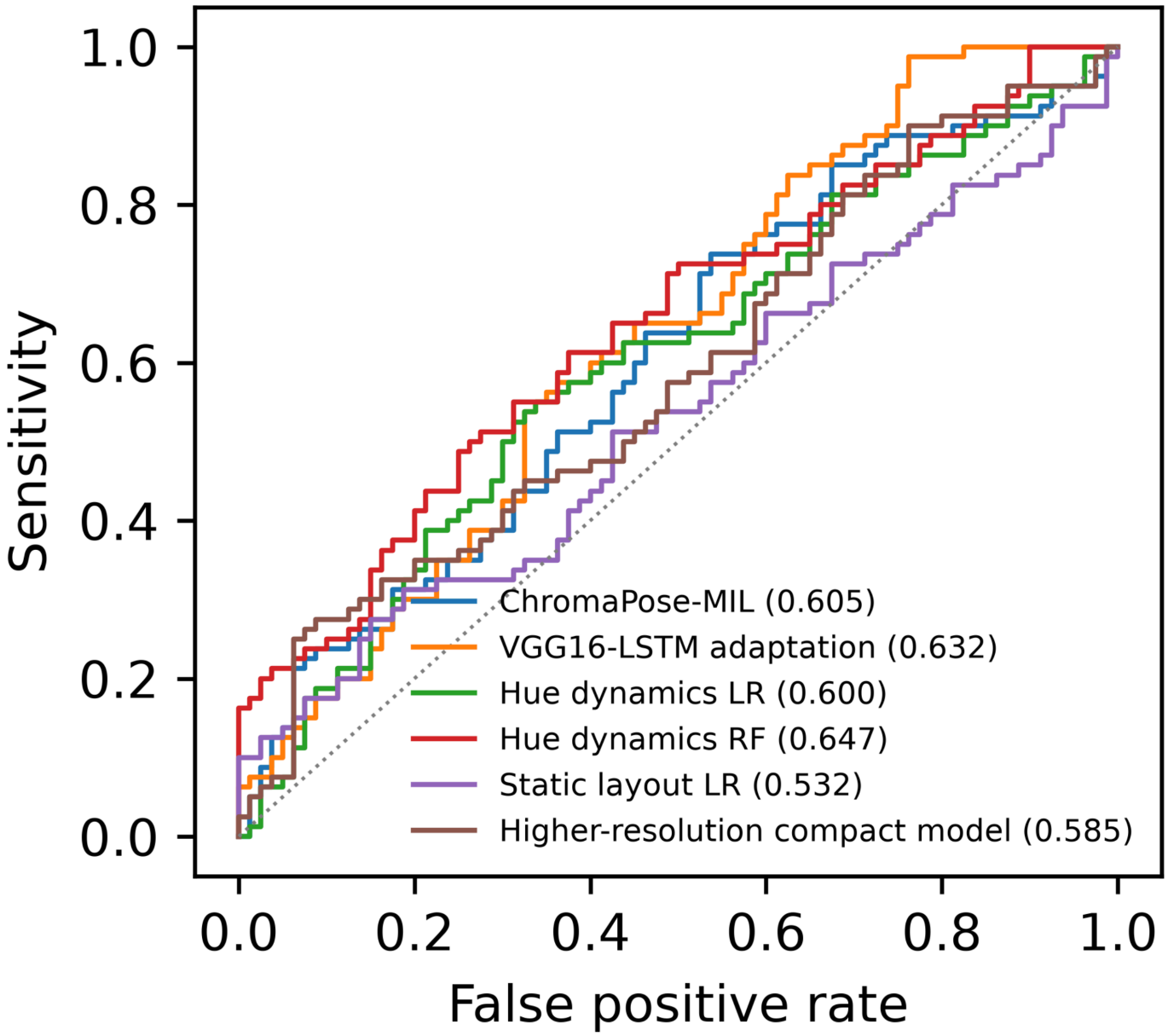


Figure 1

ROC curves from complete child-level out-of-fold predictions, seed zero. Legend entries identify only completed comparisons.

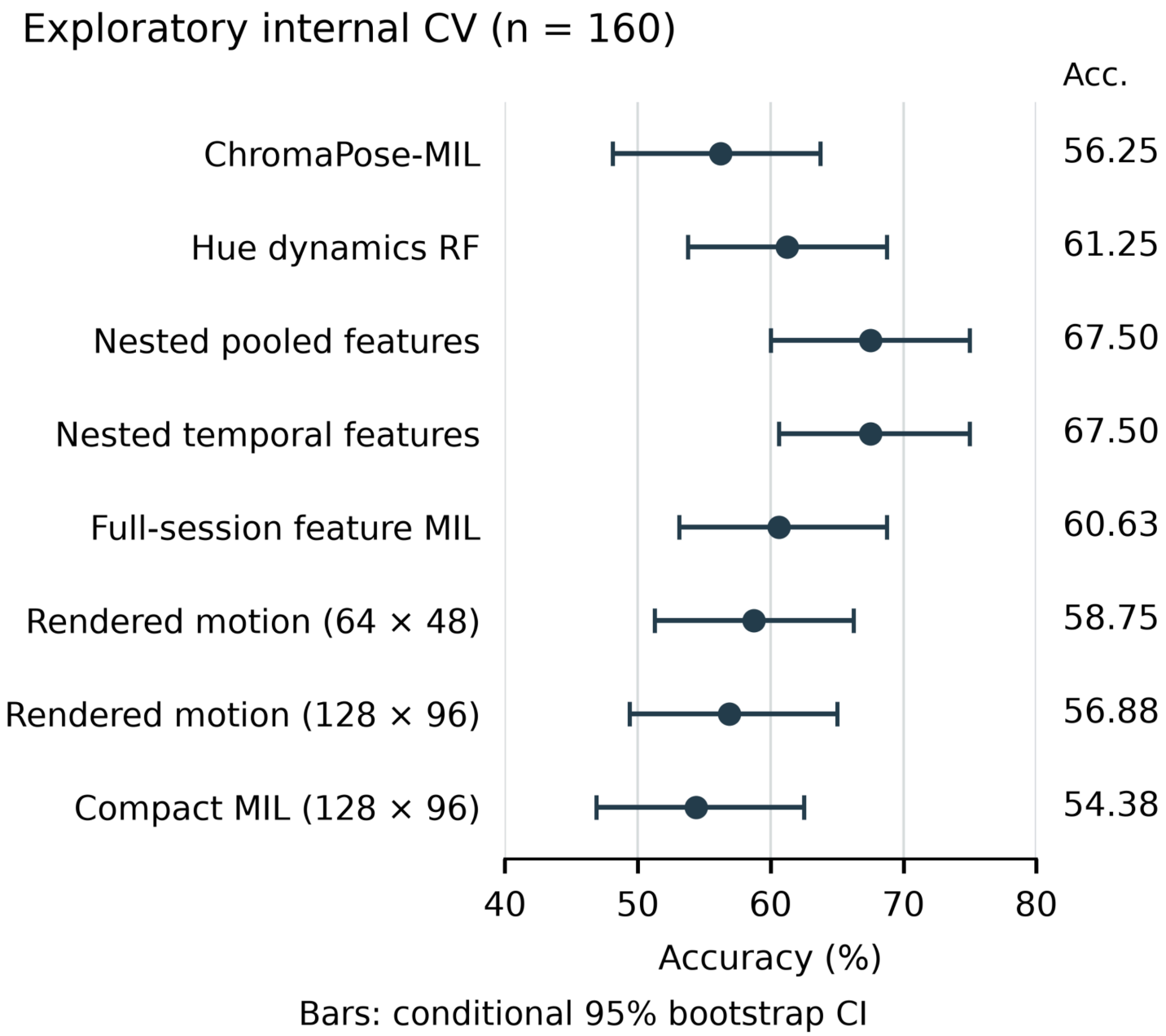


Figure 2

Accuracy with conditional 95% child-bootstrap intervals for complete methods on the same primary outer folds. These intervals condition on saved predictions and omit project-level adaptation and training variability.

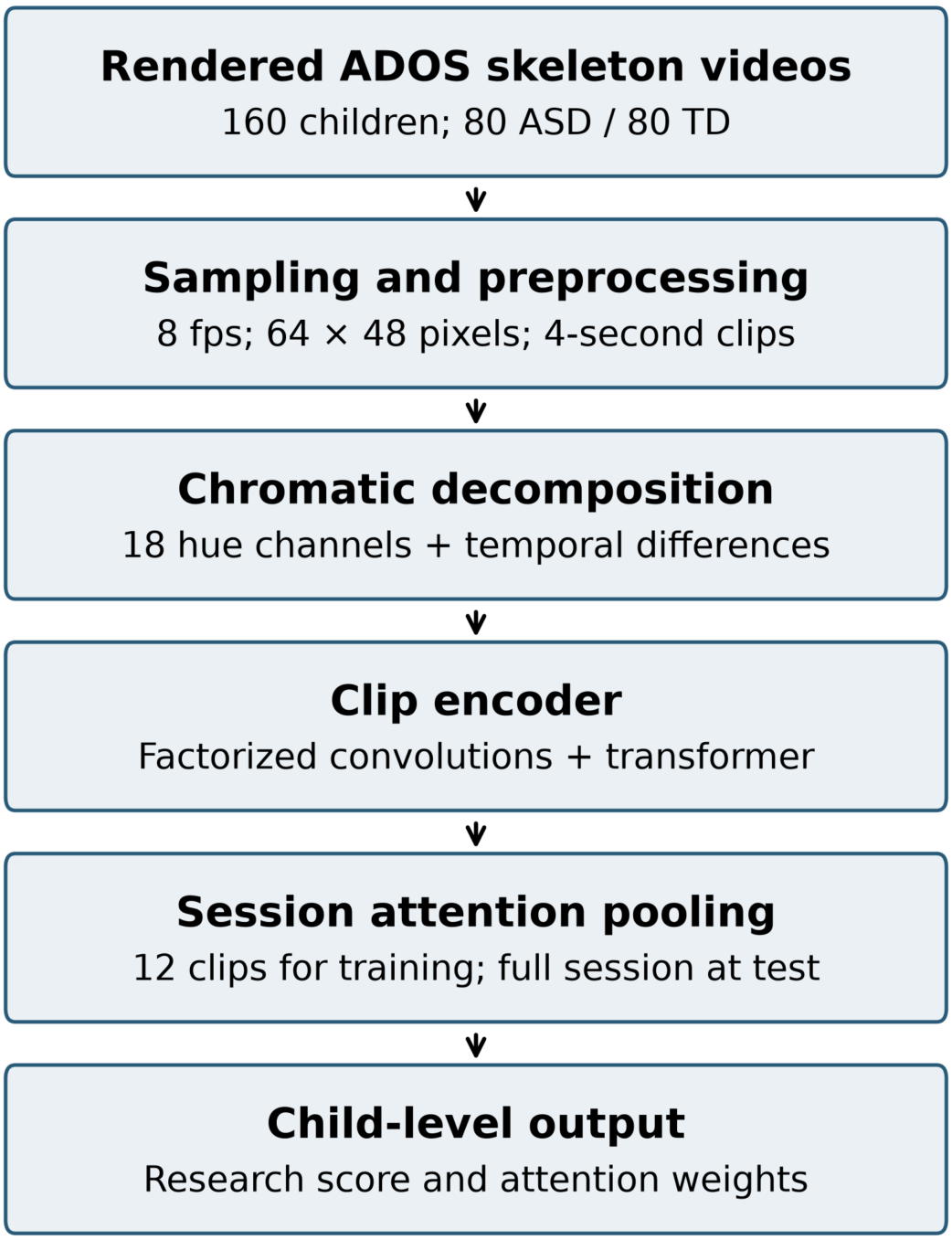


Figure 3

ChromaPose-MIL processing stages. The classifier is trained from random initialization on already-rendered skeleton videos; pose estimation itself is outside the trained pipeline.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AutismSupplementaryInformation.docx