Supplementary Materials of "Naturalistic Facial Dynamics Enable Quantitative Clinical Assessment of Atypical Expression Phenotypes in Children with Autism Spectrum Disorder

Minghao Du #, Ping Shi #, Zehao Liu, Yunuo Xu, Xiaoya Liu, Wei Liu *, Shuang Liu *, Dong Ming

* Corresponding author: lance1971@163.com; shuangliu@tju.edu.cn

Index

Supplementary Table 1. Emotion transition-probability values for the ASD group................2

Supplementary Table 2. Emotion transition-probability values for the TD group 3

Supplementary Table 3. Classification performance across activity conditions......................4

Supplementary Figure 1. Proportion of usable frames after quality control for ASD and TD groups 5

Supplementary Table 4. Head-movement deviations analysis (rad) and group comparison ...6

Supplementary Table 5. MANOVA results for effects of diagnostic group, gender, and their interaction on facial-expression features 7

Supplementary Table 1. Emotion transition-probability values for the ASD group.

From emotion	To emotion				
	Neutral	Happy	Sad	Surprise	Anger
Neutral	-	13.25%	44.90%	34.87%	6.89%
Happy	62.99%	-	12.57%	15.21%	6.16%
Sad	73.39%	5.11%	-	13.03%	8.43%
Surprise	73.92%	8.04%	16.31%	-	1.67%
Anger	40.92%	10.18%	32.75%	9.07%	-

Each table reports the average probability of transitioning from one emotion state (rows) to another (columns) across the entire session. Values on the diagonal are not included, as the duration feature separately captures the stability or maintenance of each emotion state. These matrices provide the quantitative transition patterns that complement the group-difference results reported in the main text.

Supplementary Table 2. Emotion transition-probability values for the TD group.

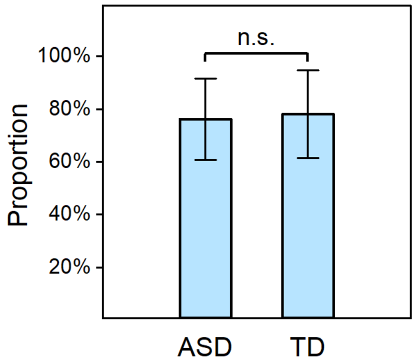
From emotion	To emotion				
	Neutral	Happy	Sad	Surprise	Anger
Neutral	-	11.16%	48.99%	36.47%	3.32%
Happy	55.98%	-	11.04%	22.46%	2.87%
Sad	80.47%	4.84%	-	12.23%	2.40%
Surprise	72.26%	8.13%	16.16%	-	1.06%
Anger	41.51%	8.70%	22.93%	10.39%	-

Each table reports the average probability of transitioning from one emotion state (rows) to another (columns) across the entire session. Values on the diagonal are not included, as the duration feature separately captures the stability or maintenance of each emotion state. These matrices provide the quantitative transition patterns that complement the group-difference results reported in the main text.

Supplementary Table 3. Classification performance across activity conditions.

	Sensitivity	Specificity	PPV	NPV	Accuracy	F1-score	AUC
Cartoon-only	0.887	0.882	0.896	0.872	0.885	0.891	0.953
Non-cartoon	0.919	0.847	0.875	0.900	0.886	0.897	0.943
Full-session	0.960	0.882	0.905	0.949	0.924	0.931	0.977

To examine whether activity type influenced classification, we manually annotated each session into cartoon-only and non-cartoon interaction segments. The non-cartoon included playing with toys, looking at picture books, and free conversation, which often overlapped and therefore were treated as one category. The same feature-extraction and classification pipeline used in the main experiment was applied separately to each activity subset. Cartoon-only segments achieved an accuracy of 88.5%, non-cartoon segments 88.6%, and the full-session data 92.4%. Performance was comparable between cartoon-only and non-cartoon segments, suggesting that the predictive features are not task-dependent. The improvement observed in the full-session condition is consistent with the advantage of longer cumulative duration rather than differences in activity content. These results indicate that classification performance remains stable across spontaneous activities, supporting our emphasis on naturalistic interaction rather than discrete and task-driven elicitation. While this comparison is reported for completeness, we do not recommend activity-based stratification in practice, as children naturally move fluidly across activities in real-world environments. The findings further highlight that naturalistic, unconstrained social interaction provides sufficient and robust behavioral signal for ASD identification.



Supplementary Figure 1. Proportion of usable frames after quality control for ASD and TD groups. The n.s. label denotes that the group comparison was not statistically significant.

In preprocessing stage, all videos underwent a unified quality control pipeline to ensure that facial-behavior features were extracted from reliable frames. Facial visibility was first screened using OpenCV's confidence-based Haar-Cascade detector, and frames without a stable face were further excluded when the emotion-recognition model produced no valid face-emotion output. For AU extraction, OpenFace confidence scores were applied to retain only frames with reliable facial-landmark tracking, helping maintain consistent facial visibility while preserving natural interaction.

To evaluate whether quality control affected the two groups differently, the proportion of usable frames was computed for each participant. A Mann-Whitney U test showed no significant group difference (ASD: 76.16±15.29%, TD: 78.07±16.49%, $p = 0.598$), indicating broadly comparable data quality across diagnostic groups. Although single-camera constraints are unavoidable, this quality control procedure represents a practical and commonly used approach for naturalistic facial-behavior studies, and multi-view recording could further enhance robustness in future work.

Supplementary Table 4. Head-movement deviations analysis (rad) and group comparison

Dimension	**ASD deviation**	**TD deviation**	**U statistic**	***P* value**
Pitch	0.073±0.101	0.060±0.072	4685	0.185
Yaw	0.065±0.224	0.077±0.193	4106	0.779
Roll	0.017±0.079	0.015±0.075	4258	0.890

To examine whether head motion differed between diagnostic groups and might influence AU extraction, we analyzed head-pose dynamics across the entire session. Although children's posture was not constrained to preserve naturalistic interaction, the camera was positioned at a stable frontal angle, and facilitators used toys, cartoons, or subtle behavioral cues to keep children naturally oriented toward the camera. To account for individual differences in initial seating orientation, we computed each child's baseline pitch, yaw, and roll from the first 20 seconds and quantified head movement as the deviation from this baseline. A 15-frame sliding window was applied to reduce frame-level noise.

Group differences in head-movement deviation were assessed using Mann-Whitney U tests across the three axes (pitch, yaw, roll). No significant group differences were observed for any head-movement dimension ($p > 0.05$). These results indicate that ASD and TD children showed comparable levels of head motion during naturalistic interaction, suggesting that systematic differences in head movement are unlikely to account for group differences in AU-based facial-expression features.

Supplementary Table 5. MANOVA results for effects of diagnostic group, gender, and their interaction on facial-expression features

Feature	Group effect (Wilks' λ)	Group *p*	Gender effect (Wilks' λ)	Gender *p*	Interaction (Wilks' λ)	Interaction *p*
Emotion variation	0.808	0.046	0.829	0.114	0.855	0.294
Expression intensity	0.277	<0.001	0.533	0.193	0.542	0.234
Facial coordination	0.112	0.041	0.136	0.151	0.168	0.431

Note. Diagnostic group showed significant effects across all three feature domains. Neither gender nor the group x gender interaction reached statistical significance.

To evaluate whether gender or group x gender interactions influenced the facial-expression features analyzed in this study, we conducted a multivariate analysis of variance (MANOVA) with diagnostic group (ASD and TD), gender, and their interaction as fixed factors. Emotion variation, expression intensity, and facial coordination were entered as dependent variables. To control for multiple comparisons, *p* values were adjusted using the false discovery rate procedure.

Diagnostic group showed significant main effects across all three feature domains, indicating reliable differences between ASD and TD children. In contrast, neither gender nor the group x gender interaction reached statistical significance for any feature. These findings indicate that gender did not explain additional variance in facial-expression dynamics beyond diagnostic status in our cohort. Overall, this analysis suggests that the proposed features robustly differentiate ASD from TD children while showing no detectable gender-related effects in the present sample. Nevertheless, it remains possible that subtle gender differences exist, and larger or more balanced samples may be required to detect such effects.