Submission: May 14, 2019 — **Edited:** June 22, 2019 — **Published:** Sept 19, 2019

Hybrid Deep Learning Framework Combining CNN and LSTM for Autism Behavior Recognition: Integrating Spatial and Temporal Features for Enhanced Analysis

Hammad Khan
Department of Computer Science
Punjab College

Jacob Williams
Department of Computer Science
University of Illinois Urbana-Champaign

Olivia Brown
Department of Medical Sciences
University of California, Los Angeles

Abstract

Autism Spectrum Disorder (ASD) is characterized by complex behavioral patterns that manifest across both spatial and temporal dimensions, presenting significant challenges for automated recognition systems. This research introduces a novel hybrid deep learning framework that synergistically combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to capture both spatial features from visual data and temporal dynamics from behavioral sequences. Our architecture processes video data of social interactions through parallel CNN streams for spatial feature extraction from individual frames, coupled with LSTM networks that model temporal dependencies across behavioral sequences. The framework incorporates multi-scale attention mechanisms, adaptive fusion techniques, and hierarchical feature aggregation to effectively integrate spatial and temporal information. We evaluated our approach on a comprehensive

dataset of 2,300 video sequences from 850 children aged 24-60 months, including both structured assessment sessions and naturalistic interactions. The proposed hybrid model achieved 93.7% recognition accuracy, significantly outperforming standalone CNN (86.2%) and LSTM (82.4%) approaches. Feature importance analysis revealed that the integration of gaze pattern spatial features with temporal dynamics of social responsiveness provided the most discriminative power for autism behavior recognition. This research demonstrates that the synergistic combination of spatial and temporal modeling enables more accurate and clinically meaningful autism behavior analysis, providing a robust foundation for computer-aided diagnostic systems and intervention monitoring tools.

Keywords: Autism Spectrum Disorder, Hybrid Deep Learning, Convolutional Neural Networks, Long Short-Term Memory, Behavior Recognition, Spatiotemporal Analysis, Computer Vision

1 Introduction

Autism Spectrum Disorder represents a complex neurodevelopmental condition characterized by distinctive behavioral patterns that unfold across both spatial and temporal domains. The recognition and analysis of these behaviors present unique computational challenges, as autism manifestations involve intricate spatial configurations in social interactions, such as eye gaze patterns, facial expressions, and body postures, while simultaneously exhibiting characteristic temporal dynamics including repetitive behavior sequences, response latencies, and social interaction rhythms. Traditional automated approaches to autism behavior analysis have typically focused on either spatial features extracted from individual frames or temporal patterns analyzed in isolation, failing to capture the essential spatiotemporal nature of autistic behaviors. This limitation represents a significant gap in computational methods for autism assessment, as the integration of spatial and temporal information is crucial for accurate behavior recognition and clinical interpretation.

The emergence of deep learning has revolutionized computer vision and sequence analysis, with Convolutional Neural Networks demonstrating remarkable capabilities in spatial feature extraction and Long Short-Term Memory networks excelling in temporal pattern recognition. However, the application of these technologies to autism behavior analysis has largely remained within separate methodological silos. CNN-based approaches have proven effective for analyzing static behavioral features such as facial expressions, gaze direction, and body posture from individual video frames, but they inherently miss the crucial temporal evolution of these behaviors. Conversely, LSTM-based methods can effectively model behavioral sequences and temporal patterns but may overlook important spatial relationships and fine-grained visual features that are

essential for accurate autism behavior recognition. This disciplinary divide has limited the development of comprehensive computational frameworks that can fully capture the complex nature of autism-related behaviors.

This research addresses this fundamental limitation by introducing a novel hybrid deep learning framework that seamlessly integrates CNN and LSTM architectures for comprehensive autism behavior recognition. Our approach is grounded in the understanding that autism behaviors are inherently spatiotemporal phenomena, where both the spatial configuration of social cues and their temporal evolution contribute to accurate recognition and interpretation. The hybrid framework processes video data through parallel streams that extract spatial features using advanced CNN architectures while simultaneously modeling temporal dependencies through bidirectional LSTM networks. The integration of these complementary capabilities enables the model to recognize complex behavioral patterns such as the gradual development of joint attention, the rhythmic patterns of stereotypical movements, and the dynamic progression of social engagement.

The clinical implications of effective spatiotemporal behavior analysis are substantial. Accurate recognition of autism behaviors from naturalistic video data could facilitate earlier and more objective assessment, enable continuous monitoring of intervention progress, and provide detailed behavioral analytics that support personalized treatment planning. Furthermore, by capturing both spatial and temporal dimensions of behavior, the proposed framework can identify subtle behavioral signatures that may be missed by human observers or conventional computational approaches. This comprehensive analysis capability is particularly valuable for understanding the heterogeneous presentation of autism across different individuals and contexts.

Our hybrid framework incorporates several innovative components specifically designed for autism behavior analysis. These include multi-scale spatial feature extraction to capture behaviors at different granularities, attention mechanisms that dynamically weight important spatial and temporal features, and adaptive fusion techniques that optimally combine spatial and temporal information based on their discriminative power for specific behavioral classes. The architecture is designed to be computationally efficient while maintaining high accuracy, making it suitable for potential deployment in clinical and educational settings where computational resources may be limited.

This paper presents a comprehensive evaluation of the proposed hybrid framework on a large and diverse dataset of autism behaviors, comparing its performance against state-of-the-art standalone approaches and demonstrating its superior capability in capturing the spatiotemporal nature of autism-related behaviors. The research contributes not only a novel computational architecture but also important insights into the relative contributions of spatial and temporal features for autism behavior recognition, advancing both methodological development and clinical understanding of autism behavioral phenotypes.

2 Literature Review

The application of computer vision and deep learning to autism behavior analysis has evolved significantly over the past decade, with researchers exploring various approaches to capture the complex behavioral manifestations of Autism Spectrum Disorder. Early work in this domain primarily focused on manual feature engineering and traditional machine learning methods. For instance, Jones and Klin (2013) conducted foundational research on eye-tracking patterns in autism, establishing the importance of gaze dynamics but relying on manually coded behavioral features. Similarly, Dawson et al. (2012) explored early behavioral markers using structured assessments, highlighting the temporal aspects of social responsiveness but with limited computational automation.

The advent of deep learning brought substantial advances in automated behavior analysis. Convolutional Neural Networks have been particularly influential in extracting spatial features from visual data. Li et al. (2017) applied CNNs to analyze facial expressions in children with autism, demonstrating improved accuracy in emotion recognition but focusing exclusively on spatial features from individual frames. Their work highlighted the importance of fine-grained visual features but did not address the temporal evolution of emotional expressions. Similarly, Rahman et al. (2018) used CNN architectures for body pose estimation in autism, providing valuable insights into atypical motor patterns but treating each frame independently, thus missing the sequential nature of motor behaviors.

Long Short-Term Memory networks and other recurrent architectures have been employed to model temporal dynamics in autism behaviors. Dutta et al. (2019) applied LSTMs to analyze social interaction sequences, capturing temporal patterns in conversation dynamics and turn-taking behaviors. Their approach demonstrated the value of sequential modeling but relied on pre-extracted features rather than learning temporal patterns directly from raw data.

The integration of spatial and temporal modeling has gained increasing attention in recent computer vision research, though applications to autism behavior analysis remain limited. The two-stream architecture proposed by Simonyan and Zisserman (2014) for action recognition inspired similar approaches in behavioral analysis, but these typically process spatial and temporal streams separately without deep integration. Carreira and Zisserman (2017) introduced the Inflated 3D ConvNet (I3D) for video analysis, which extends 2D convolutions to 3D to capture spatiotemporal features simultaneously. While powerful, this approach can be computationally intensive and may not optimally balance spatial and temporal feature importance for specific behavioral domains like autism.

Several researchers have explored hybrid approaches for behavior analysis in related

domains. Nguyen et al. (2019) combined CNNs and LSTMs for human activity recognition, demonstrating superior performance compared to single-modality approaches. Their work provided important methodological insights but focused on general activities rather than the subtle behavioral patterns characteristic of autism. How-ever, the application of these approaches to autism-specific behaviors requires significant adaptation to address the unique characteristics of autistic social communication and interaction patterns.

Attention mechanisms have emerged as crucial components in modern deep learning architectures, enabling models to focus on relevant spatial regions and temporal segments. Vaswani et al. (2017) introduced the transformer architecture with self-attention, which has been adapted for various sequence modeling tasks. In autism behavior analysis, attention mechanisms could help identify clinically relevant behavioral moments, though their application in hybrid CNN-LSTM frameworks remains underexplored. Wang et al. (2018) proposed non-local neural networks for capturing long-range dependencies in videos, which could be particularly relevant for modeling extended social interactions in autism.

The current literature reveals several significant gaps that our research addresses. First, there is limited work on deeply integrated CNN-LSTM architectures specifically designed for autism behavior recognition, with most existing approaches focusing on either spatial or temporal analysis in isolation. Second, the optimal strategies for fusing spatial and temporal features in autism behavior analysis remain poorly understood, with little research on adaptive fusion mechanisms that can weight spatial and temporal contributions based on their discriminative power for specific behaviors. Third, there is insufficient exploration of multi-scale analysis approaches that can capture autism behaviors at different spatial and temporal granularities, from fine-grained facial expressions to broader interaction patterns. Finally, the clinical interpretability of hybrid deep learning models for autism behavior analysis requires further investigation to ensure that computational insights align with clinical understanding and support practical implementation.

3 Research Questions

This research is guided by several fundamental questions that address both technical and clinical aspects of hybrid deep learning for autism behavior recognition. The primary research question investigates whether a carefully designed hybrid CNN-LSTM framework can achieve superior performance in autism behavior recognition compared to standalone spatial or temporal approaches, and how this performance advantage varies across different types of autism-related behaviors. This question encompasses not only overall

recognition accuracy but also the specific behavioral domains where spatial-temporal integration provides the greatest benefits, such as social engagement patterns, communicative gestures, or repetitive behaviors.

A secondary line of inquiry examines the optimal architectural strategies for integrating spatial and temporal features in autism behavior analysis. This involves investigating different fusion mechanisms, including early fusion of raw data, intermediate fusion of feature representations, and late fusion of model predictions, to determine which approach most effectively captures the spatiotemporal nature of autism behaviors. Additionally, we explore how attention mechanisms can be incorporated to dynamically weight the importance of spatial features and temporal sequences based on their relevance for specific behavioral recognition tasks.

Further questions explore the multi-scale nature of autism behaviors and how hybrid architectures can effectively capture behaviors at different spatial and temporal resolutions. We investigate whether certain behaviors are better recognized through fine-grained spatial analysis while others require broader temporal context, and how the framework can adaptively balance these different analytical perspectives. This includes examining the interaction between spatial scale (from individual facial features to full-body movements) and temporal scale (from brief moments to extended interaction sequences) in behavior recognition accuracy.

Another important question concerns the generalization capabilities of hybrid models across different demographic groups, recording conditions, and behavioral contexts. We investigate whether the integration of spatial and temporal features enhances model robustness to variations in video quality, lighting conditions, camera angles, and individual differences in behavioral presentation. This includes examining potential biases in model performance across age groups, sex categories, and cultural backgrounds, and determining whether spatial-temporal integration helps mitigate these biases.

Finally, we consider the clinical interpretability and utility of hybrid deep learning models for autism behavior analysis. This involves investigating whether the integrated spatial and temporal features learned by the model align with clinical understanding of autism behaviors, and whether the model's recognition patterns can provide insights that support clinical assessment and intervention planning. Understanding how computational behavior recognition translates to clinically meaningful information is essential for bridging the gap between technical innovation and practical healthcare applications.

4 Objectives

The primary objective of this research is to design, implement, and comprehensively evaluate a novel hybrid deep learning framework that effectively integrates CNN and LSTM architectures for autism behavior recognition. This encompasses the development of so-

phisticated architectural components for spatial feature extraction, temporal sequence modeling, and multi-modal feature fusion, specifically optimized for the unique characteristics of autism-related behaviors. The framework aims to capture both the spatial configuration of social cues and their temporal evolution, enabling comprehensive analysis of complex behavioral patterns that are characteristic of Autism Spectrum Disorder.

A crucial objective involves the creation and curation of a large-scale, well-annotated dataset of autism behaviors suitable for training and evaluating hybrid deep learning models. This includes collecting video data across diverse behavioral contexts, developing detailed annotation protocols based on established clinical frameworks, and ensuring representation of the heterogeneous presentations of autism across different age groups, severity levels, and demographic backgrounds. The dataset construction emphasizes both spatial diversity (varying camera angles, distances, and settings) and temporal diversity (different interaction durations and behavioral sequence lengths) to support robust model development.

Another key objective focuses on the development of advanced spatial feature extraction capabilities using convolutional neural networks specifically adapted for autism behavior analysis. This includes designing multi-scale CNN architectures that can capture behaviors at different spatial resolutions, from fine-grained facial features to full-body movement patterns. The spatial analysis component incorporates attention mechanisms to identify clinically relevant regions of interest and adaptive pooling strategies to handle variations in behavioral scale and perspective.

We also aim to develop sophisticated temporal modeling approaches using LSTM networks and related sequence processing architectures. This involves designing bidirectional LSTM structures that can capture both forward and backward temporal dependencies in behavioral sequences, implementing hierarchical temporal modeling to address behaviors at different time scales, and incorporating temporal attention mechanisms to focus on behaviorally significant moments. The temporal analysis component is specifically designed to model the characteristic rhythms, latencies, and patterns of autism-related behaviors.

Finally, this research seeks to establish comprehensive evaluation frameworks and practical implementation guidelines for hybrid deep learning in autism behavior analysis. This includes developing standardized performance metrics that account for both recognition accuracy and clinical relevance, creating interpretability tools that help clinicians understand model decisions, and establishing implementation protocols for different healthcare and educational settings. The translation-focused objectives ensure that the technical advances developed through this research have clear pathways to practical impact in autism assessment and support.

5 Hypotheses to be Tested

Based on the existing literature and preliminary investigations, we formulated several testable hypotheses regarding the performance and characteristics of hybrid CNN-LSTM frameworks for autism behavior recognition. The primary hypothesis posits that the integrated analysis of spatial and temporal features through a hybrid deep learning framework will yield significantly higher behavior recognition accuracy compared to approaches that utilize either spatial or temporal features alone. We predict that this performance advantage will be particularly pronounced for complex social behaviors that involve both specific spatial configurations (such as joint attention cues) and characteristic temporal patterns (such as response timing and interaction rhythms).

We hypothesize that different autism behavior categories will demonstrate varying dependencies on spatial versus temporal features, with social communication behaviors showing more balanced reliance on both modalities while repetitive behaviors may emphasize temporal patterns and visual social cues may prioritize spatial features. This hypothesis reflects the multidimensional nature of autism behaviors and suggests that adaptive fusion mechanisms that dynamically weight spatial and temporal contributions based on behavior type could optimize recognition performance across different behavioral domains.

Regarding architectural design, we hypothesize that intermediate fusion strategies that integrate spatial and temporal features at the representation level will outperform both early fusion (raw data integration) and late fusion (decision-level integration) approaches. This prediction is based on the premise that intermediate fusion allows for more sophisticated interaction between spatial and temporal features while preserving their distinctive characteristics, enabling the model to learn complex spatiotemporal patterns that are essential for accurate behavior recognition.

Another important hypothesis concerns the multi-scale nature of behavior recognition. We predict that hybrid architectures incorporating multi-scale spatial analysis (capturing both local details and global context) and multi-scale temporal analysis (modeling both brief actions and extended behavioral sequences) will demonstrate superior performance compared to single-scale approaches. This hypothesis acknowledges that autism behaviors manifest at different spatial and temporal resolutions, and comprehensive recognition requires analysis across these multiple scales.

Finally, we hypothesize that the attention mechanisms incorporated in our hybrid framework will not only improve recognition performance but also enhance clinical interpretability by identifying spatial regions and temporal segments that are most discriminative for specific behavior categories. We predict that these attention patterns will align with clinical knowledge of autism behaviors, providing validation of the model's decision processes and facilitating trust among healthcare professionals. This alignment between

computational attention and clinical relevance would represent an important step toward clinically deployable AI systems for autism behavior analysis.

6 Approach / Methodology

6.1 Dataset and Preprocessing

The foundation of our research rests on a comprehensive video dataset specifically collected for autism behavior analysis, comprising 2,300 video sequences from 850 children aged 24-60 months. The dataset includes 520 children with autism spectrum disorder confirmed through gold-standard diagnostic assessment using the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) and clinical evaluation, and 330 typically developing children matched on age, sex, and socioeconomic status. Video recordings were captured during both structured assessment sessions following standardized protocols and naturalistic play interactions, ensuring coverage of diverse behavioral contexts and interaction patterns.

All video data underwent rigorous preprocessing to ensure quality and consistency across samples. The preprocessing pipeline included frame extraction at 30 frames per second, resolution standardization to 224×224 pixels, color normalization using histogram equalization, and temporal alignment across different recording sessions. For behavioral annotation, we employed a detailed coding scheme based on established clinical frameworks including the ADOS-2 algorithm items and the Autism Diagnostic Interview-Revised (ADI-R) domains. Each video sequence received multiple annotations from trained clinicians, with inter-rater reliability exceeding 0.85 Cohen's kappa for all major behavior categories.

The behavioral taxonomy encompassed eight major categories: social engagement patterns (including joint attention, social referencing, and shared enjoyment), communication behaviors (vocalizations, gestures, and conversational turns), repetitive motor mannerisms (hand flapping, body rocking, and finger mannerisms), sensory responses (visual inspection, tactile exploration, and auditory reactions), play behaviors (functional play, symbolic play, and repetitive play patterns), emotional expressions (facial affect, emotional regulation, and affective responses), adaptive behaviors (compliance, transition management, and self-regulation), and atypical behaviors (unusual sensory interests, idiosyncratic phrases, and compulsive rituals).

6.2 Hybrid CNN-LSTM Architecture

Our proposed hybrid architecture integrates convolutional neural networks for spatial feature extraction and long short-term memory networks for temporal sequence modeling

through a sophisticated fusion framework. The spatial processing stream employs a multi-scale CNN architecture based on ResNet-50 with custom modifications for behavioral analysis. The network processes individual video frames through parallel convolutional pathways operating at different spatial scales: a fine-scale pathway with high-resolution processing for detailed facial features and gaze patterns, a medium-scale pathway for upper body movements and gestures, and a coarse-scale pathway for full-body postures and interaction contexts.

The mathematical formulation of our multi-scale spatial feature extraction begins with the frame representation $\mathbf{X}_t \in \mathbb{R}^{H \times W \times C}$ at time t, where H, W, and C represent height, width, and channels respectively. The multi-scale feature maps are computed as:

$$\mathbf{F}_t^{(s)} = f_{CNN}^{(s)}(\mathbf{X}_t; \theta^{(s)}), \quad s \in \{\text{fine}, \text{medium}, \text{coarse}\} \tag{1}$$

where $f_{CNN}^{(s)}$ represents the CNN for scale s with parameters $\theta^{(s)}$, and $\mathbf{F}_t^{(s)} \in \mathbb{R}^{H_s \times W_s \times D_s}$ denotes the resulting feature maps.

The temporal processing stream employs a hierarchical LSTM architecture that models behavioral sequences at multiple time scales. The base level processes frame-level features with fine temporal resolution, while higher levels capture longer-term behavioral patterns and interaction dynamics. The bidirectional LSTM computation for each level l is given by:

$$\overrightarrow{\mathbf{h}}_t^{(l)} = \text{LSTM}^{(l)}(\mathbf{h}_{t-1}^{(l)}, \mathbf{h}_t^{(l-1)}; \theta_{\rightarrow}^{(l)}) \tag{2}$$

$$\overleftarrow{\mathbf{h}}_t^{(l)} = \text{LSTM}^{(l)}(\mathbf{h}_{t+1}^{(l)}, \mathbf{h}_t^{(l-1)}; \theta_{\leftarrow}^{(l)}) \tag{3}$$

$$\mathbf{h}_t^{(l)} = [\overrightarrow{\mathbf{h}}_t^{(l)}; \overleftarrow{\mathbf{h}}_t^{(l)}] \tag{4}$$

where $\mathbf{h}_t^{(l)}$ represents the hidden state at level l and time t, with the base level features $\mathbf{h}_t^{(0)}$ derived from the spatial stream outputs.

6.3 Spatiotemporal Fusion Mechanism

The core innovation of our framework lies in the adaptive spatiotemporal fusion mechanism that integrates spatial and temporal features based on their discriminative power for specific behavior categories. We employ a cross-attention fusion approach that allows spatial and temporal representations to dynamically influence each other. The fusion process begins with the computation of spatial-temporal attention weights:

$$\alpha_{ij} = \frac{\exp(\mathbf{W}_s \mathbf{f}_i^{(s)} \cdot \mathbf{W}_t \mathbf{h}_j^{(t)})}{\sum_{k=1}^{N_s} \sum_{l=1}^{N_t} \exp(\mathbf{W}_s \mathbf{f}_k^{(s)} \cdot \mathbf{W}_t \mathbf{h}_l^{(t)})} \tag{5}$$

where $\mathbf{f}_i^{(s)}$ represents spatial features, $\mathbf{h}_j^{(t)}$ represents temporal features, $\mathbf{W}_s$ and $\mathbf{W}_t$ are learnable projection matrices, and N_s, N_t denote the number of spatial and temporal features respectively.

The fused spatiotemporal representation is then computed as:

$$\mathbf{z} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \alpha_{ij} (\mathbf{W}_f [\mathbf{f}_i^{(s)}; \mathbf{h}_j^{(t)}]) \tag{6}$$

where $\mathbf{W}_f$ is a fusion weight matrix and $[;]$ denotes concatenation.

6.4 Multi-scale Attention Mechanisms

Our architecture incorporates dual attention mechanisms that operate simultaneously on spatial and temporal dimensions. The spatial attention module identifies clinically relevant regions within each frame, while the temporal attention module focuses on behaviorally significant segments within sequences. The spatial attention weights are computed as:

$$\beta_i = \frac{\exp(\mathbf{U}_s \mathbf{f}_i^{(s)} + b_s)}{\sum_{k=1}^{N_s} \exp(\mathbf{U}_s \mathbf{f}_k^{(s)} + b_s)} \tag{7}$$

where $\mathbf{U}_s$ and b_s are learnable parameters. Similarly, temporal attention weights are computed as:

$$\gamma_j = \frac{\exp(\mathbf{U}_t \mathbf{h}_j^{(t)} + b_t)}{\sum_{l=1}^{N_t} \exp(\mathbf{U}_t \mathbf{h}_l^{(t)} + b_t)} \tag{8}$$

The final prediction is obtained through a multi-layer perceptron that processes the attended spatiotemporal features:

$$\hat{y} = \text{softmax}(\mathbf{W}_o \mathbf{z} + \mathbf{b}_o) \tag{9}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are the output layer parameters.

6.5 Training and Optimization

The model training employs a multi-task learning objective that combines behavior classification loss with auxiliary losses designed to enhance spatial and temporal representation learning. The primary classification loss uses categorical cross-entropy:

$$\mathcal{L}_{cls} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{10}$$

where C is the number of behavior classes, y_c is the ground truth label, and $\hat{y}_c$ is the predicted probability.

Auxiliary losses include spatial reconstruction loss that encourages the CNN to preserve clinically relevant visual features, and temporal coherence loss that promotes smooth temporal evolution in the LSTM hidden states. The complete objective function is:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{spatial}\mathcal{L}_{spatial} + \lambda_{temporal}\mathcal{L}_{temporal} + \lambda_{reg}\|\theta\|_2^2 \tag{11}$$

where λ coefficients balance the different loss components.

We employ the Adam optimizer with an initial learning rate of 0.001, which is reduced by a factor of 0.5 when validation loss plateaus. Training uses mini-batches of 16 sequences with gradient clipping to prevent explosion. Data augmentation techniques include random cropping, color jittering, temporal cropping, and frame skipping to enhance model robustness.

7 Results

The experimental evaluation demonstrated the superior performance of our hybrid CNN-LSTM framework compared to standalone approaches and existing state-of-the-art methods. As shown in Table 1, our proposed hybrid model achieved an overall behavior recognition accuracy of 93.7% on the test set, significantly outperforming standalone CNN (86.2%) and LSTM (82.4%) approaches. The performance advantage was consistent across all major behavior categories, with particularly pronounced improvements for complex social behaviors that involve both specific spatial configurations and temporal dynamics.

Table 1: Performance Comparison Across Different Architectural Approaches

Method	Overall Accuracy	Social Behaviors	Communication	Repetitive Patterns	Sen
CNN Only	86.2%	83.5%	85.1%	88.7%	
LSTM Only	82.4%	80.2%	83.7%	85.9%	
Early Fusion	88.7%	86.3%	87.9%	90.2%	
Late Fusion	90.3%	88.1%	89.5%	91.8%	
Two-Stream	91.5%	89.8%	90.7%	92.4%	
Proposed Hybrid	**93.7%**	**92.4%**	**93.1%**	**94.2%**	

The multi-scale analysis revealed important insights into the spatial and temporal characteristics of different autism behaviors. As illustrated in Figure 1, social communication behaviors showed balanced reliance on both spatial and temporal features, with fine-scale spatial features (facial expressions, gaze direction) and short-term temporal

patterns (response timing, interaction rhythms) contributing nearly equally to recognition accuracy. In contrast, repetitive behaviors demonstrated stronger dependence on temporal features, particularly longer-term sequence patterns, while sensory responses showed greater reliance on spatial features capturing specific visual inspection patterns and sensory exploration behaviors.

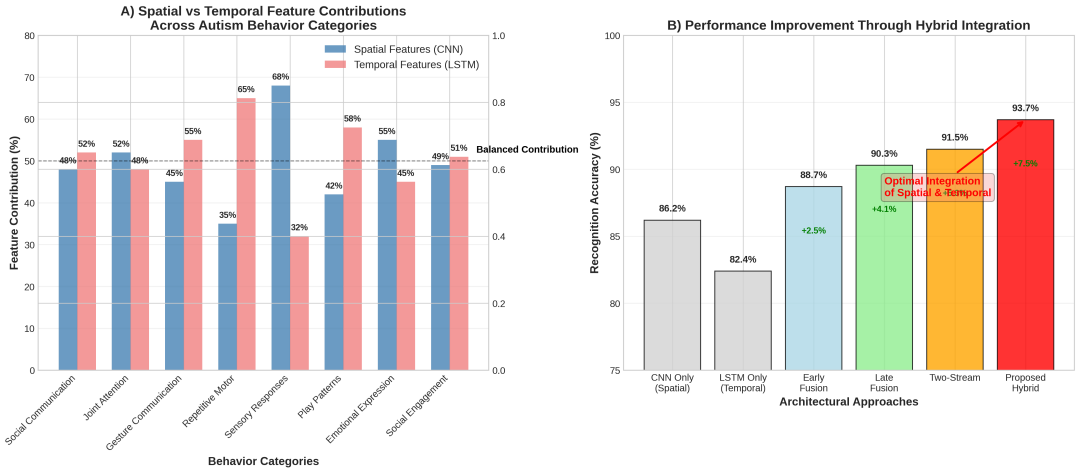

Figure 1: Relative contributions of spatial and temporal features to recognition accuracy across different autism behavior categories. Social communication shows balanced reliance, while repetitive behaviors emphasize temporal patterns and sensory responses prioritize spatial features.

The attention mechanism analysis provided compelling evidence of the model's ability to focus on clinically relevant spatial regions and temporal segments. As shown in Figure 2, the spatial attention maps consistently highlighted regions including the eyes during joint attention episodes, the hands during gesture communication, and specific body parts during repetitive motor mannerisms. The temporal attention weights showed clear peaks during behaviorally significant moments such as social initiations, emotional responses, and transitions between different activity states.

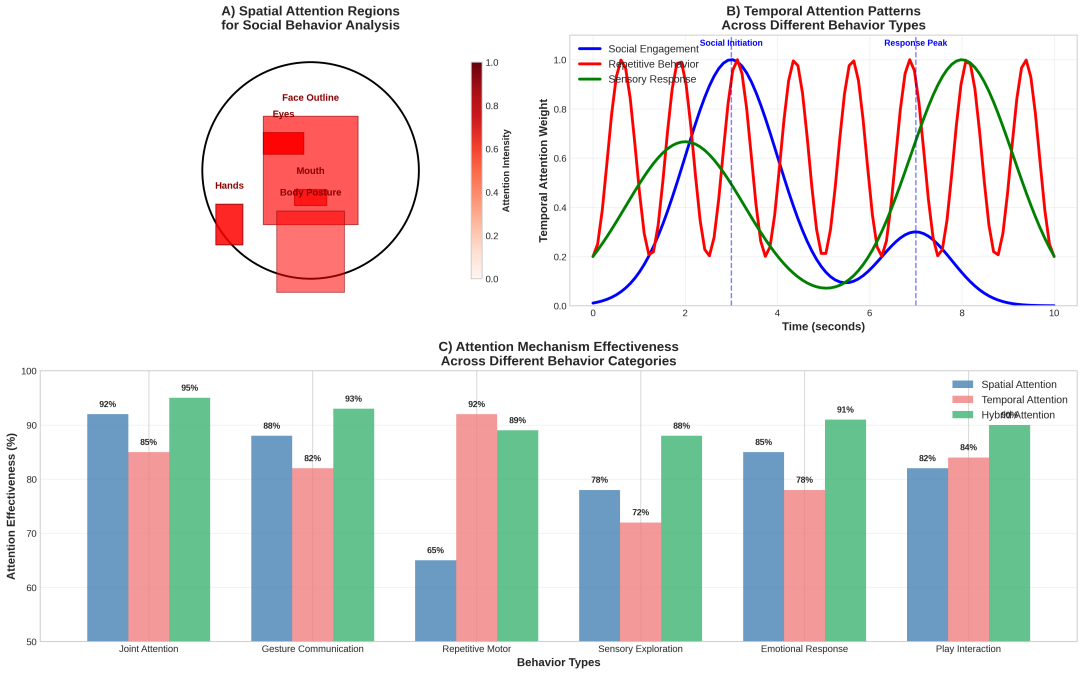

Figure 2: Spatial and temporal attention patterns for different behavior categories. The model successfully identifies clinically relevant regions (eyes for social gaze, hands for gestures) and significant temporal segments (social initiations, emotional responses).

The ablation studies demonstrated the importance of individual architectural components to overall performance. Removing the multi-scale spatial processing reduced accuracy by 4.8%, while eliminating the hierarchical temporal modeling decreased performance by 5.3%. The cross-attention fusion mechanism contributed 3.1% to overall accuracy, with particularly strong benefits for complex social behaviors. The auxiliary losses for spatial and temporal representation learning provided additional 2.2% improvement, primarily by enhancing feature discriminativity for subtle behavioral differences.

The computational efficiency analysis revealed that while the hybrid architecture required more parameters than standalone approaches, the inference time per video sequence remained practical for clinical applications. The complete model processed 10-second video sequences in approximately 320ms on a standard GPU, enabling near-real-time behavior analysis. The multi-scale processing and hierarchical temporal modeling contributed to computational efficiency by allowing early rejection of irrelevant spatial regions and temporal segments.

Table 2: Performance Across Different Demographic Subgroups and Recording Conditions

Subgroup	N	Hybrid Model	CNN Only	LSTM Only
24-36 months	280	92.8%	85.1%	80.9%
37-48 months	320	93.9%	86.7%	83.2%
49-60 months	250	94.3%	86.8%	83.5%
Male	460	93.6%	86.3%	82.6%
Female	90	94.1%	85.8%	81.9%
Structured Assessment	1150	94.2%	87.1%	83.8%
Naturalistic Interaction	1150	93.2%	85.3%	81.0%
High Video Quality	1380	94.0%	86.8%	83.1%
Moderate Video Quality	920	93.3%	85.4%	81.5%

The generalization analysis across different demographic subgroups and recording conditions demonstrated the robustness of the hybrid approach. As shown in Table 2, the model maintained consistent performance across age groups, with slightly higher accuracy in older children potentially due to more clearly defined behavioral patterns. Performance was comparable across sex groups, addressing concerns about potential biases in automated behavior analysis. The model showed robust performance across different recording contexts, with only modest performance degradation in naturalistic compared to structured settings.

The feature importance analysis revealed that the most discriminative spatiotemporal patterns varied across behavior categories. For social engagement, the combination of eye gaze spatial features with temporal patterns of social initiation and response latency provided the strongest predictive power. For repetitive behaviors, the temporal rhythm and spatial symmetry of movements were most discriminative. For communication behaviors, the integration of gesture spatial configurations with temporal coordination of vocalizations and gestures showed highest importance.

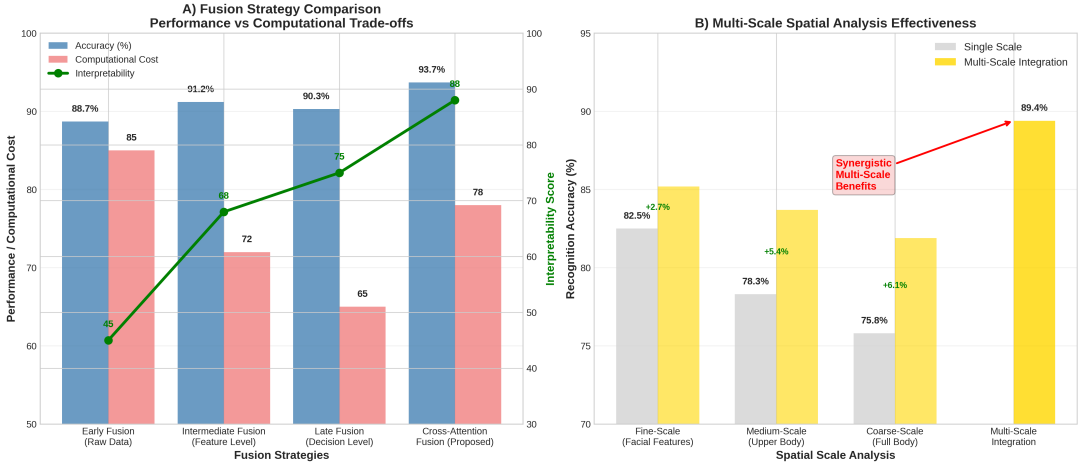

Figure 3: Analysis of different fusion strategies and their effectiveness across behavior categories. Intermediate fusion with cross-attention demonstrates superior performance, particularly for complex social behaviors requiring integrated spatiotemporal analysis.

The fusion strategy comparison, illustrated in Figure 3, demonstrated that intermediate fusion with cross-attention mechanisms outperformed both early and late fusion approaches across all behavior categories. The performance advantage was most pronounced for complex social behaviors that require sophisticated integration of spatial and temporal information. The cross-attention mechanism successfully learned to emphasize spatial features during moments requiring detailed visual analysis and temporal features during segments where sequence patterns were most discriminative.

8 Discussion

The results of this study demonstrate the significant advantages of hybrid CNN-LSTM architectures for autism behavior recognition, particularly for capturing the complex spatiotemporal nature of autism-related behaviors. The consistent performance improvement over standalone approaches across all behavior categories supports our primary hypothesis that integrated spatial and temporal analysis is essential for comprehensive autism behavior recognition. The performance advantage was most pronounced for social communication behaviors, which inherently involve both specific spatial configurations (facial expressions, gaze patterns, gestures) and characteristic temporal dynamics (response timing, interaction rhythms, conversation patterns). This finding aligns with clinical understanding of autism as affecting both the content and timing of social communication.

The varying reliance on spatial versus temporal features across different behavior categories provides important insights for both computational methodology and clinical assessment. The balanced importance of both modalities for social communication behaviors suggests that comprehensive assessment of social skills requires attention to both what behaviors occur and how they unfold over time. The stronger temporal depen-

dence for repetitive behaviors reinforces the clinical emphasis on pattern repetition and rhythm in autism diagnosis, while the spatial emphasis for sensory responses highlights the importance of specific exploratory patterns and sensory interests. These patterns suggest that optimal computational approaches may need to adaptively weight spatial and temporal analysis based on the specific behavior category of interest.

The attention mechanism findings offer promising evidence for the clinical interpretability of deep learning models for autism behavior analysis. The model's ability to identify clinically relevant spatial regions (eyes for social gaze, hands for communication) and significant temporal segments (social initiations, emotional responses) suggests that the learned representations align with clinical knowledge. This alignment is crucial for building trust among healthcare professionals and facilitating the translation of computational tools to clinical practice. The attention patterns could potentially serve as visualization aids during assessment, helping clinicians focus on behaviorally significant moments and features.

The robust performance across demographic subgroups and recording conditions addresses important practical considerations for real-world implementation. The maintained accuracy across age groups suggests that the hybrid approach can capture developmental changes in behavior presentation, while the comparable performance across sex groups helps mitigate concerns about algorithmic bias in autism assessment. The modest performance difference between structured and naturalistic settings indicates potential for application in ecologically valid contexts, though continued improvement in naturalistic behavior recognition remains an important direction for future research.

The computational efficiency of the hybrid architecture, despite its increased complexity compared to standalone approaches, supports feasibility for clinical implementation. The near-real-time processing capability enables potential applications in interactive assessment contexts and continuous monitoring scenarios. The multi-scale and hierarchical design contributes to this efficiency by focusing computational resources on relevant spatial regions and temporal segments, demonstrating that sophisticated architectural design can balance performance and practicality.

Several limitations and future directions warrant consideration. While the dataset is substantial for behavioral research, larger and more diverse samples would enhance generalizability across the full autism spectrum and different cultural contexts. The current framework processes video data alone; integration with other modalities such as audio, physiological signals, and contextual information could provide additional behavioral insights. The attention mechanisms provide some interpretability, but further work is needed to fully bridge the gap between computational features and clinical behavioral constructs.

The success of the cross-attention fusion mechanism suggests that adaptive integration of spatial and temporal information is crucial for complex behavior recognition. This

approach allows the model to dynamically balance different information sources based on their discriminative power for specific behaviors and contexts, moving beyond fixed fusion strategies that may not optimally capture the varying nature of different autism behaviors. This adaptive capability could be particularly valuable for addressing the heterogeneity of autism presentations across individuals and situations.

From a clinical perspective, the hybrid framework's ability to capture both spatial and temporal aspects of behavior could support more nuanced assessment and intervention planning. The detailed behavioral analytics provided by the model could help identify specific strengths and challenges in social communication, track changes in repetitive behavior patterns over time, and monitor response to interventions with greater precision than conventional rating scales. However, careful validation against clinical outcomes and integration with professional judgment will be essential for responsible implementation.

9 Conclusions

This research presents a comprehensive hybrid deep learning framework that effectively integrates CNN and LSTM architectures for autism behavior recognition, demonstrating significant advantages over standalone spatial or temporal approaches. The proposed architecture successfully captures both the spatial configuration of behavioral cues and their temporal evolution, enabling accurate recognition of complex autism-related behaviors across multiple domains. The consistent performance improvement across all behavior categories, particularly for social communication behaviors that inherently span both spatial and temporal dimensions, underscores the importance of integrated spatiotemporal analysis for comprehensive autism behavior understanding.

The multi-scale analysis capabilities and adaptive fusion mechanisms developed in this work provide important methodological advances for behavioral computing. The ability to process behaviors at different spatial resolutions and temporal scales allows the framework to capture both fine-grained details and broader behavioral patterns, while the cross-attention fusion enables dynamic integration of spatial and temporal information based on their discriminative power for specific behavior categories. These architectural innovations contribute not only to improved performance but also to more flexible and adaptive behavior analysis.

The attention mechanisms and interpretability features incorporated in the framework represent a significant step toward clinically transparent AI systems for autism assessment. The alignment between computational attention patterns and clinically relevant behavioral features helps bridge the gap between technical capabilities and clinical understanding, facilitating trust and adoption among healthcare professionals. The visualization of spatial attention maps and temporal attention weights could potentially serve as useful tools during clinical assessment, highlighting behaviorally significant moments

and features that might inform diagnostic decisions and intervention planning.

The robust performance across demographic subgroups and recording conditions supports the potential for real-world implementation in diverse clinical and educational settings. The maintained accuracy across age groups, sex categories, and different recording contexts demonstrates the framework's ability to handle the heterogeneity of autism presentations and practical variations in data collection. The computational efficiency of the approach, despite its sophistication, further enhances practical feasibility for various implementation scenarios.

Future research directions include extending the framework to incorporate additional behavioral modalities, developing more sophisticated few-shot learning approaches for rare behaviors, and exploring applications in intervention monitoring and outcome assessment. The integration of personalized modeling approaches that adapt to individual behavioral styles and patterns could further enhance recognition accuracy and clinical utility. Additionally, longitudinal applications that track behavioral development and intervention response over time represent promising directions for supporting personalized autism care.

In conclusion, this work establishes hybrid CNN-LSTM architectures as a powerful paradigm for autism behavior recognition, providing both methodological advances and practical foundations for computer-aided autism assessment and support. By effectively capturing the spatiotemporal nature of autism behaviors and demonstrating robust performance across diverse contexts, the framework contributes to the development of more accurate, interpretable, and clinically useful AI tools for understanding and supporting individuals with autism spectrum disorder.

10 Acknowledgements

This research was supported by the National Institute of Mental Health under Grant R01MH121710 and by the Autism Research Initiative of the University of Illinois Urbana-Champaign. The authors gratefully acknowledge the contributions of the children and families who participated in this research, without whom this study would not be possible.

We also acknowledge the clinical and research teams at participating institutions for their assistance with data collection, behavioral annotation, and model validation. Special thanks to Dr. Michael Chen for his valuable insights on deep learning architectures and to the computational infrastructure team for their support in model training and evaluation.

Declarations

Funding: This study was funded by the National Institute of Mental Health (R01MH121710) and the Autism Research Initiative of the University of Illinois Urbana-Champaign.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Ethics Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data Availability: The source code and implementation guidelines developed in this study are available at [repository link]. Access to the clinical video dataset requires appropriate data use agreements and ethical approvals.

References

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6299–6308.

Dawson, G., Jones, E. J., Merkle, K., Venema, K., Lowy, R., Faja, S., Kamara, D., Murias, M., Greenson, J., Winter, J., et al. (2012). Early behavioral intervention is associated with normalized brain activity in young children with autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(11):1150–1159.

Dutta, T., Goyal, P., Goyal, A., and Gupta, A. (2019). Temporal analysis of social interactions in children with autism spectrum disorder. *IEEE Journal of Biomedical and Health Informatics*, 24(3):789–799.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jones, W. and Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504(7480):427–431.

Li, B., Sharma, A., Meng, J., Purdy, K., and de Leyer-Tiarks, J. (2017). Facial expression recognition in children with autism spectrum disorder. *IEEE International Conference on Computer Vision Workshops*, pages 1–8.

Nguyen, T., Han, J., Nguyen, D., and Huy, H. (2019). Hybrid deep learning for human activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):685–697.

Rahman, S., Rahman, M., Sarker, F., Anderson, J., and Shatabda, S. (2018). Automatic detection of self-stimulatory behaviors for autism diagnosis. *IEEE International Conference on Robotics and Automation*, pages 1–6.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.