Spatiotemporal Transformers with Multiple Instance Learning for Label-Efficient Behavioral Analysis in Autism (Student Abstract)

Emily Yu

Massachusetts Institute of Technology
emily_yu@mit.edu

Abstract

The identification of unique traits and behavior is essential to providing personalized intervention in individuals with Autism Spectrum Disorder. However, the limited personalized quantitative data with experts' annotations in autism research pose a fundamental challenge to train AI models for unique behavioral patten discovery. Multiple Instance Learning (MIL) has demonstrated promising results in medical domains, where annotations are only needed at the group level (i.e., a whole sequence) instead of individual data instances. It provides a cost-effective way to train statistical models with limited labeled data. Additionally, the rise of pretrained models have shown great success in improving the performance in few-shot learning scenarios. In this proof-of-concept study, we propose a novel framework that integrates a transformer encoder pre-trained on large-scale spatiotemporal data with MIL, for unique behavioral pattern detection from autistic individuals. Our results demonstrated the discrimination of individual-level autistic behavioral differences and the accurate classification of behaviors across distinct groups: typically developing (TD) and autistic (ASD). Beyond aggregate performance metrics, we highlight visual insights from temporal instance scores, revealing interpretable differences between individuals in their respective groups. These results show promising progress towards tools that can be used for personalized intervention for autistic individuals, and more interpretable AI diagnostics.

Source Code — github.com/mathjams/AAAI26

Introduction

Autism Spectrum Disorder (ASD) is a neurological developmental disorder that affects around 1 in 31 individuals in the United States (Centers for Disease Control and Prevention (CDC) 2025). Autistic individuals commonly display characteristic behaviors, such as repetitive body motions or abnormal responses to stimuli. These behaviors are often coping mechanisms, with responses that are unique to the individual. Meanwhile, the advancement of AI, especially deep learning techniques, have shown promising progress in autistic intervention. Additionally, the collection of sensory input through virtual reality (VR) systems provide quantitative information about the behavioral patterns of autistic individuals, allowing for tailored treatment and support. This is especially important in autistic intervention, due to the large diversity of the spectrum.

However, the effectiveness of these approaches heavily relies on the access to plentiful labeled data sources, essential for training complex models capable of discovering spatiotemporal patterns from time-series behavioral sequences. Collecting large quantities of temporal datasets with detailed annotations that pinpoint when and where unique behavioral patterns occur is challenging and often requires specialized domain knowledge. Although methods such as data augmentation or data synthesis provide effective means to address data scarcity, these methods still face major limitations, requiring significant research from both domain experts and the machine learning community. Additionally, datasets on specific groups of autistic individuals are especially rare. Thus, understanding the common characteristics of spatiotemporal patterns is critical to detect the unique traits embedded in the behavioral sequences of autistic individuals.

To address the lack of annotated temporal behavior data and limited dataset sizes, we propose a novel framework that seamlessly integrates two complementary learning components. First, we utilize PatchTST, a transformer model pretrained on large-scale public time series data. It helps to encode common spatiotemporal features from eye fixation sequences from autistic and TD adolescents collected through a VR tracking system. These features are then fed into a Multiple Instance Learning (MIL) model, which relies only on the group-level supervision (i.e., whether a behavioral sequence is collected from an autistic or TD individual) to detect unique behavioral patterns to best differentiate these two groups. MIL has demonstrated success in many medical domains, including autism research, where fully annotated datasets are often unavailable (Carbonneau et al. 2018). Built upon the pre-trained time-series transformer, the MIL model can fully leverage temporal behavioral data to effectively detect fine-grained behavioral patterns (i.e., subsequences of eye-fixations in a short time interval) that are uniquely associated with autistic individuals. By integrating these two complimentary learning components, the proposed framework allows us to leverage the benefits of large-scale models while working with smaller datasets.

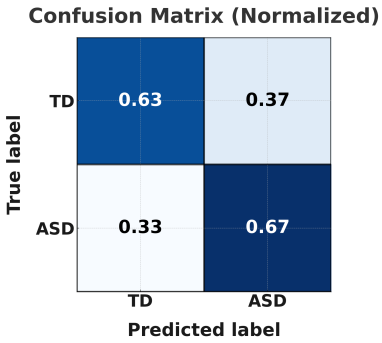

Figure 1: Confusion matrix demonstrating balanced class predictions

Methodology

Data were collected in earlier work using the Multimodal Virtual Classroom Interface (MVCI) (Yu et al. 2023). In this setup, the VR system continuously tracked eye gaze coordinates as participants completed tasks displayed on the screen. Each gaze location is represented by an (x, y) coordinate, where $x, y \in (0, 1)$. From these recordings, sequences of eye fixations were extracted. The MVCI allowed participants to complete tasks that would be performed in a classroom environment, in particular, copying an image from the reference panel onto another section of the screen as shown in Figure 2. The screen contains a reference panel for selecting the color, a main task region for completing the task, and the background with avatars (providing visual and audio stimuli), as shown in Appendix A (Yu 2026).

Since only sequence-level diagnostics labels are available from our data collection, we employ Multiple Instance Learning (MIL) as a weakly supervised approach (Carbonneau et al. 2018). In this setting, negative bags (TD) contain only negative instances, while positive bags (ASD) include at least one positive instance. Each instance includes a fixed number of consecutive fixations within the sequence, and all the instances form a bag. Thus, *positive instances can be interpreted as containing the distinctive behaviors patterns differentiating autistic individuals from their typically developing peers*.

We segmented each fixation sequence into fixed-length overlapping windows, each representing an instance of behavior. We extracted temporal features through a pretrained PatchTST encoder (Nie et al. 2022), and computed instance scores through a lightweight MIL head. Instance embeddings were aggregated via top-k pooling at the bag level. Training was optimized with a weighted hinge loss to address label imbalance. Performance was evaluated by a weighted F1 score for early stopping, with the confusion matrix for visualization. Additional details are included in Appendix B (Yu 2026).

Experimental Results

We trained the models using 75% of the data and tested on the rest. To expand our training and validation sets, we also generated subsequences from the original dataset. The confusion matrix in Figure 1 shows the balanced classification performance across ASD and TD groups (with recalls of 0.67 and 0.63, respectively), which indicates that our model does not classify in favor one group over the other despite imbalanced sample sizes, essential in rare pattern detection.

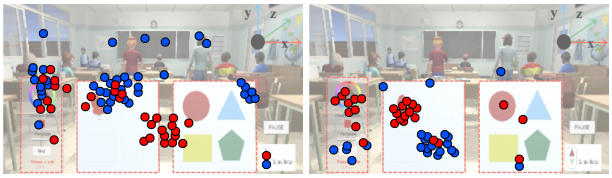

Figure 2: Color-coded eye fixations for Individuals 1 and 2

Case study. Now we analyze two specific autistic individuals, and with gaze fixations colored red and blue, where red points indicate fixations from positively labeled instances and blue points indicate fixations from negatively labeled instances. For `Individual 1` as shown in Figure 2, there is a cluster of abnormal fixations located between the two task areas. This pattern suggests that `Individual 1` may have unique behaviors when transitioning between tasks, specifically when shifting from viewing the reference image to reproducing it. Moreover, we see that the fixations that lie on the background are not labeled as positive instance points, indicating that response to background stimuli are unlikely to explain the observed behavioral differences.

For `Individual 2` in Figure 2, we see that there are concentrated fixations around the reference and main task panels. This distribution could indicate potential difficulties with task mechanics, such as replicating the figure or selecting the correct colors, reflected by the positively labeled fixations near the color selection panel. Additional results and visualizations are included in Appendix C (Yu 2026).

Discussion. Our results demonstrate that the proposed framework can effectively capture discriminatory fixations patterns by only using sequence level labels. The balanced accuracy verifies that it differentiates between the two classes through behavioral abnormalities and avoids class bias. It also demonstrates the ability of the PatchTST encoder to extract meaningful spatiotemporal features, enabling the lightweight MIL head to separate classes with limited training data. Our findings highlight that the framework uncovers heterogeneity in autistic behaviors, where positive behavioral instances are **different** between autistic individuals. Thus, this framework can effectively detect fine-grained differences at both the group and subject levels.

Conclusion

By integrating a pre-trained time-series transformer with a weakly supervised MIL model, the proposed framework effectively extracts unique behavioral patterns from limited behavioral data with only sequence level labels. It alleviates the reliance on fine-grained behavioral annotations, offering a more practical and cost-effective way to support personal intervention and treatment in autism research.

Acknowledgments

I would to thank Dr. Zhi Zheng from University of Notre Dame, who served as my mentor during the course of working on this project. Dr. Zheng is an expert in human-computer interaction with years of experience in designing reliable assistive systems for mental health care. Dr. Zheng introduced to me this research area, suggested the overall research topic of using a data-driven approach to understand unique behaviors in children with autism, and shared the dataset that is analyzed in my study. We had weekly meetings to discuss the progress of my work and Dr. Zheng offered important feedback that helped to shape this research. I would also like to thank Zhiwei Yu, who helped preprocess the data by extracting the fixations from the raw data.

References

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353.

Centers for Disease Control and Prevention (CDC). 2025. Data and Statistics on Autism Spectrum Disorder. https://www.cdc.gov/autism/data-research/index.html. Accessed: 2025-09-02.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Yu, E. 2026. Appendix: Spatiotemporal Transformers with Multiple Instance Learning for Label Efficient Behavioral Analysis in Autism. https://github.com/mathjams/AAAI2026.

Yu, Z.; Iadarola, S.; Daley, S.; and Zheng, Z. 2023. A Multimodal Virtual Classroom Interface to Facilitate Discovery of Behavioral Patterns in Response to Sensory Stimuli. *International Society for Autism Research Annual Meeting 2023 (INSAR 2023)*.