Deep Temporal Sets with Evidential Reinforced Attentions for Unique Behavioral Pattern Discovery

Dingrong Wang [1] **Deep Shankar Pandey** [1] **Krishna Prasad Neupane** [1] **Zhiwei Yu** [1]
Ervine Zheng [1] **Zhi Zheng** [1] **Qi Yu** [1]

Abstract

Machine learning-driven human behavior analysis is gaining attention in behavioral/mental healthcare, due to its potential to identify behavioral patterns that cannot be recognized by traditional assessments. Real-life applications, such as digital behavioral biomarker identification, often require the discovery of complex spatiotemporal patterns in multimodal data, which is largely under-explored. To fill this gap, we propose a novel model that integrates uniquely designed Deep Temporal Sets (DTS) with Evidential Reinforced Attentions (ERA). DTS captures complex temporal relationships in the input and generates a set-based representation, while ERA captures the policy network's uncertainty and conducts evidence-aware exploration to locate attentive regions in behavioral data. Using child-computer interaction data as a testing platform, we demonstrate the effectiveness of DTS-ERA in differentiating children with Autism Spectrum Disorder and typically developing children based on sequential multimodal visual and touch behaviors. Comparisons with baseline methods show that our model achieves superior performance and has the potential to provide objective, quantitative, and precise analysis of complex human behaviors.

1. Introduction

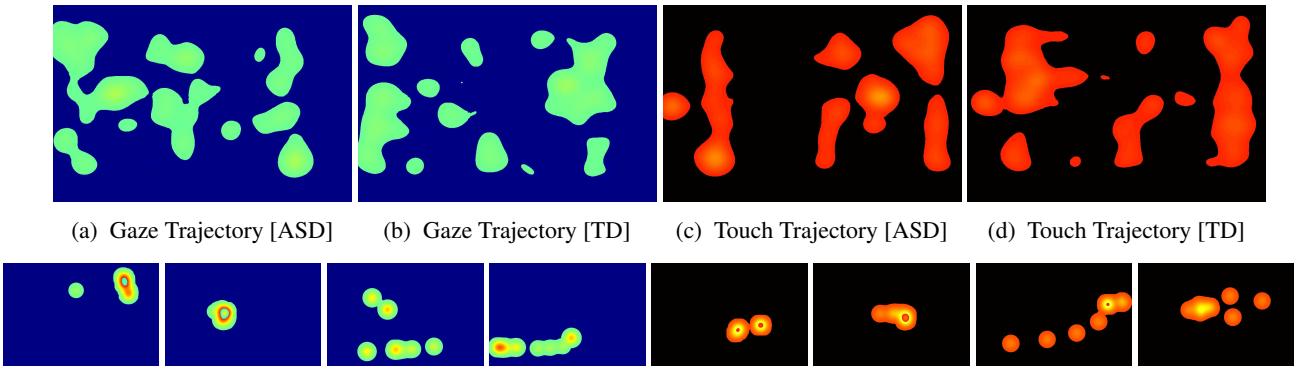
Machine learning has been widely applied to detecting and analyzing human behaviors (Chen et al., 2021), with recent advances in behavioral/mental health care (Thieme et al., 2020). Digital technologies enable the collection of big and high-resolution data, which further empower machine learning and artificial intelligence in the automated recognition of medical conditions and treatment planning (Washington et al., 2020). However, there are still important bottlenecks. Meaningful observations of patients lead to multimodal data where complex spatiotemporal patterns are hidden and hard to catch (Cai et al., 2019). Collections of data may be incomplete and noisy, and thus lead to erroneous predictions (Le Glaz et al., 2021). In addition, health research requires a higher level of interpretability on the analytical results, making many existing models (*e.g.,* deep neural networks) unsuitable since their black-box nature cannot provide insights about the data.

To address these challenges, we propose novel integration of uniquely designed **D**eep **T**emporal **S**ets with **E**vidential **R**einforced **A**ttentions (DTS-ERA) to identify signature behavioral patterns (SBPs) based on multimodal behavioral dynamics. To evaluate the performance of the proposed model, we applied it to analyze complex spatiotemporal datasets collected from a series of computer games designed to identify the unique visual and touch behavioral patterns in response to sensory stimuli for children with Autism spectrum disorder (ASD). ASD is a prevalent (1 in 100 worldwide (Zeidan et al., 2022) and 1 in 44 in the U.S. (Maenner et al., 2021)) developmental disorder characterized by challenges in social communications as well as restricted and repetitive behaviors (APA, 2013). Human-computer interaction systems have been developed to study responsive behaviors through controlled stimuli delivery and high-resolution behavioral data collection (Bozgeyikli et al., 2017; Koirala et al., 2021), which generates important but complex spatiotemporal patterns hiding in an overwhelming amount of data points (Noel et al., 2017). In addition, users may not attend to the whole interaction, causing noises and incomplete data. Existing sequential models that process the step-wise time series data may fail to achieve satisfactory performance and extract interpretive embeddings facing these unique challenges, because the informative features and noises are weighted equally.

In the new DTS-ERA model, the DTS first extracts temporal features from the raw behavioral data and transforms the step-wise observations into a series of latent representations. Such representations can capture the sequential dependen-

[1]Rochester Institute of Technology. Correspondence to: Qi Yu <qi.yu@rit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.



(a) Gaze Trajectory [ASD] (b) Gaze Trajectory [TD] (c) Touch Trajectory [ASD] (d) Touch Trajectory [TD]

(e) ISS [ASD] (f) Circle [ASD] (g) GS [TD] (h) LL [TD] (i) BCS [ASD] (j) Slide [ASD] (k) GL [TD] (l) LS [TD]

Figure 1: (a)-(d): Heatmap visualization of one ASD (ID:2) and one TD (ID:3) users' gaze and touch trajectories; (e)-(h): signature behavioral patterns (SBPs) discovered from ASD and TD gaze trajectories, where ISS, GS, and LL stands for imbalanced scatter switching, global scan, and local line, respectively; (i)-(l) SBPs discovered from ASD and TD touch trajectories, where BCS, GL, and LS stand for balanced cluster switching, global line, and local scan, respectively.

cies manifested in the various patterns of users' behaviors, which is the key to analyzing temporal data. Unlike conventional sequential models, which tend to forget early time steps, the DTS-generated representation properly balances the early and recent time steps, which helps identify unusual patterns if they happen early. We further integrate DTS with evidential reinforced attentions (ERA) to perform evidence-aware exploration over DTS embeddings and attentively select and build sub-spatiotemporal sets. In particular, we introduce a uniquely designed reward function to encourage attention to the unknown but potentially important behavioral sub-sequences (called signature behavioral patterns or SBPs). The reward simultaneously considers both the prediction accuracy for exploitation and evidence-based uncertainty estimation for unknown behavior exploration. Inspired by the task formulation in few-shot learning works, we design the training objective as a sub-trajectory classification problem. DTS-ERA takes sub-trajectory data for training rather than the entire trajectory and thus is capable of processing incomplete sequential data, which is common for behavioral studies of children. As a result, informative behavioral patterns can be effectively identified (with noisy ones excluded) to ensure good interpretability along with improved predictive accuracy.

We use illustrative examples, as shown in Figure 1, to provide snapshots of experimental results and demonstrate the effectiveness of the model design. As can be seen, it is very hard to distinguish ASD and typically developing (TD) children by simply comparing their entire gaze and touch (*i.e.,* painting ball movement) trajectories as shown in Figures 1a-1d, where brightness indicates gaze duration and touch hardness, respectively. This is because the whole trajectory contains both noisy behaviors and common patterns between autistic children and their TD peers. On the other hand, Figures 1e-1l show a set of representative SBPs discovered from the ASD and TD gaze and touch trajectories, respectively by the proposed DTS-ERA model. They show clear distinctions between the ASD and TD groups of children, which are evidenced by the thorough statistical analysis (see Table 2) conducted as part of our experiments. Meanwhile, the SBPs are highly interpretable, which unveils important behavioral differences between the two groups of users. For example, Autistic users tended to stare at fixed locations for a longer time and push or lift the painting ball with a touch sensation harder while swiping the ball less. In contrast, TD users looked around a wider area across the game board, and moved the ball in a relatively linear way, probably by following the designed painting path. Furthermore, by placing attention to these SBPs in the DTS embedding, it achieves a highly impressive classification accuracy, demonstrating the potential of using the model to facilitate behavioral phenotyping. We summarize our main contributions below:

- a novel end-to-end DTS-ERA model to analyze sparse, multimodal, dynamic, and noisy behavioral data.
- deep temporal sets to generate spatio-temporal set encoding, capable of capturing special behavioral patterns.
- evidential reinforced attentions to identify SBPs that are both discriminative and highly interpretable,
- use of task formulation to achieve accurate predictions with limited sparse information from incomplete sequences of behavioral data, making it more realistic and effective to support real-world behavioral studies.

2. Related Work

Machine learning driven digital behavioral biomarker discovery. In recent years, there has been a growing interest in identifying data-driven biomarkers leveraging machine learning techniques (Babrak et al., 2019; Bent et al., 2021;

Lee et al., 2021). These biomarkers have unique advantages over traditional biomarkers such as analysis at both the individual and population level, continuous measures, and passive monitoring (Babrak et al., 2019). Lee et al. (Lee et al., 2021) leverage various machine learning approaches with putative biomarkers as an imbalance between sympathetic and parasympathetic nervous activity to predict cognitive fatigue. Further, establishing robust neuroimaging biomarkers using structural magnetic resonance imaging (MRI) with traditional machine learning mechanisms are used to diagnose and tailor treatment for ASD patients (Pagnozzi et al., 2018). In contrast to these approaches, our model is uniquely designed to capture users' signature behavioral patterns to better differentiate different user groups.

Deep sets. Deep Sets (DS) are a novel class of models that operate on sets (Zaheer et al., 2017). DS can handle varying input lengths and be applied to a wide range of downstream applications, including classification (Gordon et al., 2018; Gondal et al., 2021) and regression (Garnelo et al., 2018b;a). Attentive Neural Processes (ANP) (Kim et al., 2019) includes an attention-based encoder-decoder architecture and ANP-RNN (Qin et al., 2019) further extends ANP model with an RNN-based encoder structure to handle regression problems with temporal dependencies in the input. In contrast to the self-attention or cross-attention from ANP-RNN, the proposed DTS-ERA introduces novel evidential reinforced attention to select observations and build sub-spatiotemporal sets from multimodal gaze-touch trajectories. This mechanism allows the model to identify the representative SBPs in response to sensory stimuli and helps the model more accurately differentiate between the behaviors of children with ASD and those of TD.

Reinforcement learning. RL has been increasingly used to solve computer vision and natural language processing problems. For example, (Mnih et al., 2014) applies reinforced visual attention to recognize important image patches for digit classification. (Paulus et al., 2017) introduces a neural network model with novel intra-attention and a new training method that combines standard supervised word prediction and RL. In medical assessment, (Ye et al., 2020) proposes an RL-based synthetic sample selection method that learns to choose synthetic images containing reliable and informative features.

Our work designs a new reward function that balances classification accuracy and evidence-based exploration. Instead of performing relatively simple synthetic sample selection, we combine RL and Deep Temporal Sets (DTS) in novel ways to achieve evidential reinforced attentions to handle complex and sparse sequential multimodal data.

3. Preliminaries

Data collection. The data used for this article were collected

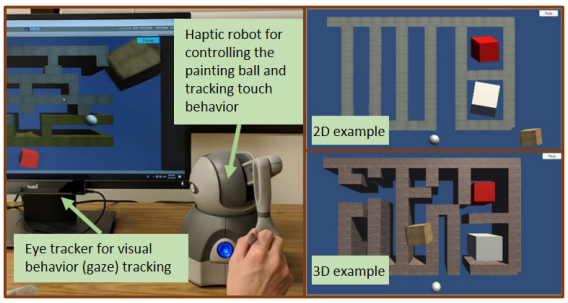


Figure 2: SAVR setup used for data collection

using multiple virtual reality (VR) games (Koirala et al., 2021). While all gaming data were collected using a similar setup, where participants sat in front of a screen-based VR, the contents were quite different, including 2D and 3D `Maze Painting`, `Word Scanning`, and `Coloring` following a template. Figure 2 shows an example of the interface for collecting different modality data, in our case, gaze and touch, as well as `Maze Painting` 2D and 3D environment. `Word Scanning` and `Coloring` games were presented in a VR classroom showing on the screen, shown in Figure 13 of Appendix E.3. Additional details can be found in Appendix B.

Evidential learning and uncertainty. Evidential learning is an evidence acquisition process where every training sample adds support to learn higher order evidential distribution (Sensoy et al., 2018; Amini et al., 2020). Given the target y_n, is drawn i.i.d. from a Gaussian distribution with unknown mean and variance (μ, σ^2) the model evidence can be introduced by further placing a prior distribution on (μ, σ^2). Leveraging Gaussian prior on the unknown mean and the Inverse-Gamma prior on the unknown variance, the posterior of (μ, σ^2) is the Normal-Inverse-Gamma (NIG) distribution. Given a NIG posterior distribution, we can derive the mean ($\mathbb{E}[\mu]$), aleatoric ($\mathbb{E}[\sigma^2]$) and epistemic (Var$[\mu]$) uncertainty as:

$$\mathbb{E}[\mu] = \gamma, \ \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}, \ \text{Var}[\mu] = \frac{\beta}{\nu(\alpha - 1)} \quad (1)$$

4. Deep Temporal Sets with Evidential Reinforced Attentions

Overview. The proposed DTS-ERA model seamlessly integrates deep temporal sets with evidential reinforced attentions for signature behavioral pattern discovery from dynamic multi-modal sensory inputs. Figure 3 presents a high-level overview of the model. The DTS module extracts temporal features from raw input data and transforms them into a series of latent representations organized into a set. Then, state representations are generated by aggregating the DTS-generated embeddings while paying special attention

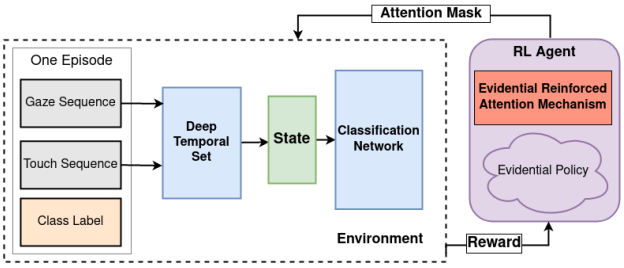


Figure 3: Overview of the proposed DTS-ERA

to certain entries in the set, where the attended subset is generated by an RL agent leveraging the ERA module. ERA performs evidence-aware exploration of the DTS-generated embeddings through a specially designed reward that simultaneously considers both the prediction accuracy for exploitation and evidence-based uncertainty estimation for unknown behavior exploration.

4.1. Model Design

We aim to develop a model $\mathcal{F}$ that can accurately predict, and identify the SBPs from multimodal sequential data. In this work, we focus on SBPs that can be used to effectively distinguish ASD and TD children given the recorded multimodal behavioral observations (*i.e.,* gaze and touch) during the game-play:

$$\mathcal{F} : \{\mathbf{g}_n, \mathbf{t}_n\}_{n=1}^{N_e} \rightarrow y;\ \mathbf{g}_n \in \mathbb{R}^{M_g}, \mathbf{t}_n \in \mathbb{R}^{M_t} \quad (2)$$

where $y \in [0, 1], \mathcal{T} = \{\mathbf{g}_n, \mathbf{t}_n\}_{n=1}^{N_e}$ represents the entire data within an episode (*a.k.a.,* trajectory). In this trajectory, $(\mathbf{g}_n, \mathbf{t}_n)$ represents n^{th} instance and N_e is the number of gaze and touch data points in the episode. Each gaze feature is M_g dimensional, each touch feature is M_t dimensional, $y = 1$ represents an ASD user, and $y = 0$ is a TD user. The length of trajectory N_e varies across the users and episodes.

Inspired by the task formulation in few-shot learning (Garnelo et al., 2018a; Gordon et al., 2018), we design the training objective as a sub-trajectory classification problem. Specifically, we randomly sample a sub-trajectory $\mathcal{T}^s = \{\mathbf{g}_n, \mathbf{t}_n\}_{n=k}^{k+N_s}$ of length N_s ($\forall k \in [1, N_e - N_s]$) from the trajectory $\mathcal{T}$ (ignoring padding $3p$ for simplicity) and train the model to accurately identify the user group based on the limited sub-trajectory information (*i.e.,* $\mathcal{F} : \mathcal{T}^s \rightarrow y$). This addresses the limited data problem, enables the model to train on a large number of training tasks, and encourages the model to capture multiple identifying patterns of users. Moreover, this sub-trajectory-based classification is likely to be more realistic and representative in real-world settings especially involving children. Simply, children are likely to paint a maze in multiple rounds and in each round, focus on painting for a couple of seconds (a sub-trajectory or sequence) instead of a single round of painting. Some children may not even complete the game and we may only have partial trajectory information available.

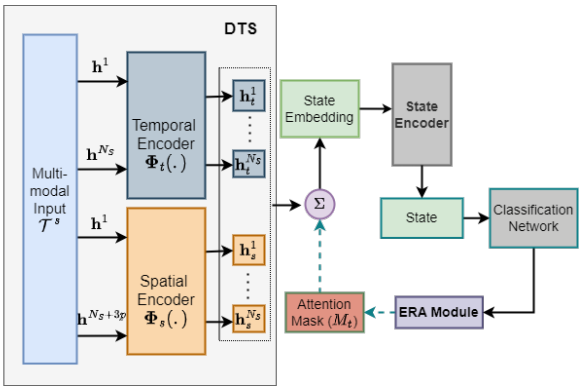


Figure 4: Architecture of Deep Temporal Sets (DTS)

Deep Temporal Sets (DTS). DTS, as shown in Figure 4, consists of a temporal encoder Φ_t which encodes each embedding using all past sequence information and a spatial encoder Φ_s which generates embedding by considering nearby local information in the sequence. Note that we apply separate sequence encoders for different modalities (*e.g.,* gaze and touch) in the implementation to fully leverage the multimodal information. The multimodal inputs are passed through temporal (*e.g.,* LSTM or GRU networks) and spatial encoders (*e.g.,* FCN) to capture the past and local information from the sequence and generate the encoding $\mathbf{h}_t^k$ and $\mathbf{h}_s^k$, respectively, $\forall k \in [1, N_s]$, where each encoding matches with the original instance in the input sequence $\mathcal{T}^s$. We instantiate the spatial encoder with the FCN network using three Convolution-1D blocks with kernel size p, thus requiring a longer sequence with size $N_s + 3p$ as input by considering future p instances as padding in each block. Each encoding can be seen as the representation for the multimodal sequential input around time step k. Then, those deep-set encodings output by different sequence encoders are concatenated and aggregated through an average pooling operation to obtain the deep-set encoding $\mathbf{d}$:

$$\begin{aligned}\mathbf{d} &= \texttt{DTS}([\mathbf{h}^1, ..., \mathbf{h}^{N_s}]) \\ &= \frac{1}{N_s}\sum_{i=1}^{N_s} \texttt{concat}[\Phi_t(\mathbf{h}^i), \Phi_s(\mathbf{h}^i)]\end{aligned} \quad (3)$$

Here, `concat` is a concatenation function. The deep-set encoding operation for the multimodal sequential inputs is permutation invariant, *i.e.,* for any permutation P:

$$\texttt{DTS}([\mathbf{h}^1, ..., \mathbf{h}^{N_s}]) = \texttt{DTS}([\mathbf{h}^{P(1)}, ..., \mathbf{h}^{P(N_s)}]) \quad (4)$$

This design can be seen as a multimodal realization of deep sets encoder of (Zaheer et al., 2017) that takes the form $f(X) = \rho(\sum_{x \in X} \phi(x))$ where ϕ is the sequence encoder, and ρ is the mean aggregation function.

Evidential Reinforced Attentions (ERA). In ERA (shown in Figure 5), we first construct the current state embedding

($\mathbf{e}_t$) by concatenating the DTS-generated embedding $\mathbf{d}$ and the RL-agent selected attentive subset embedding $\mathbf{d}^t_{attn}$:

$$\mathbf{e}_t = \texttt{concat}(\mathbf{d}, \mathbf{d}^t_{attn}) \tag{5}$$

We leverage a state encoder (SE), which takes the current state embedding ($\mathbf{e}_t$) and previous state ($\mathbf{s}_{t-1}$) to generate current state-space ($\mathbf{s}_t$) as:

$$\mathbf{s}_t = \texttt{SE}(\mathbf{e}_t, \mathbf{s}_{t-1}; \theta_{se}) \tag{6}$$

We develop an evidential policy network (π_{θ_e}) parameterized by θ_e, which takes the $\mathbf{s}_t$ as an input and output evidential distribution parameters ($\gamma, \nu, \alpha, \beta$), where the meanings of these parameters are given in Section 3. Then, the likelihood of choosing an action, $\mathbf{a}_t$ is obtained by marginalizing over the prior parameters (μ, σ^2):

$$\begin{aligned} &p(\mathbf{a}_t|\gamma, \nu, \alpha, \beta) \\ &= \int_{\sigma^2}\int_{\mu} p(\mathbf{a}_t|\mu, \sigma^2)p(\mu, \sigma^2|\gamma, \nu, \alpha, \beta)\mathrm{d}\mu\mathrm{d}\sigma^2 \\ &= \mathrm{St}\left(\mathbf{a}_t; \gamma, \beta(1+\nu)/(\nu\alpha), 2\alpha\right) \end{aligned} \tag{7}$$

where $\mathrm{St}(\mathbf{a}_t; \gamma, \beta(1+\nu)/(\nu\alpha), 2\alpha)$ is the Student-t distribution with location, scale, and degrees of freedom respectively, which is achieved by placing a NIG evidential prior on a Gaussian likelihood.

From this Student-t distribution, we sample an action $\mathbf{a}_t$ that provides the attentive location in DTS and directs the policy update, respectively. It should be noted that our action is a continuous vector of length N_a representing all the starting position of an attention window. Specifically, for each attention $a^k_t (k \in [1, N_a])$, we apply a sigmoid function (σ) and multiply it with length of $(N_s - W)$, where W represents the size of attention window. We then generate the attention starting index using a floor function:

$$idx^k_t = \lfloor \sigma(a^k_t) \cdot (N_s - W) \rfloor \tag{8}$$

We construct an all-zero mask M_t of length N_s and then flip the entries indexing in $[idx^k_t, idx^k_t + W], \forall k \in [1, N_a]$ to 1. The RL selected attentive subset embedding in time step t $\mathbf{d}^t_{attn}$ is then obtained by $M_t \cdot \texttt{concat}(\Phi_t(\mathbf{h}^i), \Phi_s(\mathbf{h}^i))$, where $\cdot$ symbolizes dot product function.

We design a novel evidential reward function that incorporates standard RL reward computed with predicted classification accuracy and epistemic uncertainty that captures policy network's uncertainty while providing an action:

$$\begin{aligned} r^e(\mathbf{s}_t, \mathbf{a}_t) &= r(\mathbf{s}_t, \mathbf{a}_t) + \lambda \texttt{epistemic}(\pi_{\theta_e}(\cdot|\mathbf{s}_t)) \\ r(\mathbf{s}_t, \mathbf{a}_t) &= \mathbb{1}\{p_T = y_s\} \end{aligned} \tag{9}$$

where $r(\mathbf{s}_t, \mathbf{a}_t)$ is a predictive reward representing the classification accuracy at last time step T, p_T is the last time step's predicted result while y_s is the user category label corresponding to sub trajectory $\mathcal{T}^s$, and $\texttt{epistemic}(\pi_{\theta_e}(\cdot|\mathbf{s}_t)) = \mathrm{Var}[\mu] = \frac{\beta}{\nu(\alpha-1)}$ is the epistemic uncertainty.

Given the evidential reward, we introduce an epistemic value function, $V^e(\mathbf{s}_t)$, which can be computed by repeatedly applying the Bellman operator (B^π):

$$B^\pi V^e(\mathbf{s}_t) \triangleq r^e(\mathbf{s}_t, \mathbf{a}_t) + \gamma_{RL}\mathbb{E}_{\mathbf{s}_{t+1}\sim\pi}[V(\mathbf{s}_{t+1})] \tag{10}$$

The detailed workflow of the ERA module is presented in Figure 5. The module takes $\mathbf{s}_t$ as input to the evidential policy network that generates evidential distribution parameters. We further marginalize those parameters and achieve a predictive student-t distribution and from which we sample an action $\mathbf{a}_t$. We generate attention masks based on the provided action and then select attentive gaze and touch embeddings and compute evidential reward simultaneously.

4.2. Derivation of Epistemic Policy Iteration

We derive epistemic policy iteration to achieve optimal policy by alternating between epistemic policy evaluation and epistemic policy improvement.

Lemma 1 (Epistemic Policy Evaluation)**.** *Given the Bellman operator B^π in (10) and $V^{n+1} = B^\pi V^n$, the value will converge to the epistemic value of policy π as $n \to \infty$.*

Lemma 2 (Epistemic Policy Improvement)**.** *Given a new policy π_{new} that is updated via (15), then $V^e_{\pi_{new}}(\mathbf{s}_t) \geq V^e_{\pi_{old}}\mathbf{s}_t$ for all $\mathbf{s}_t$.*

Theorem 3 (Epistemic Policy Iteration)**.** *Alternating between epistemic policy evaluation and epistemic policy improvement for any policy $\pi \in \Pi$ converges to an optimum epistemic policy π^* such that $V^{\pi^*}(\mathbf{s}_t) \geq V^e_\pi(\mathbf{s}_t)$ for all $\mathbf{s}_t$.*

Please refer to Appendices D.1, D.2 and D.3 for proofs.

4.3. Training and Inference

The training procedure involves the parameter update associated with both DTS and ERA modules. The DTS module is trained with supervised learning utilizing binary cross-entropy loss as:

$$\begin{aligned} &\mathcal{L}(\theta_d, \theta_{NN}) \\ &= -\frac{1}{T}\sum_{t=1}^{T} y_s \cdot \log(\texttt{NN}(\mathbf{s}_t)) + (1 - y_s) \cdot \log(1 - \texttt{NN}(\mathbf{s}_t)) \end{aligned} \tag{11}$$

where NN is a binary classifier neural network, and y_s is a ground truth label. Then the DTS module is updated as:

$$\theta_d \longleftarrow \theta_d - \eta_d \nabla_{\theta_d}\mathcal{L}(\theta_d, \theta_{NN}) \tag{12}$$

Also, the classification network is updated as:

$$\theta_{NN} \longleftarrow \theta_{NN} - \eta_{NN} \nabla_{\theta_{NN}}\mathcal{L}(\theta_d, \theta_{NN}) \tag{13}$$

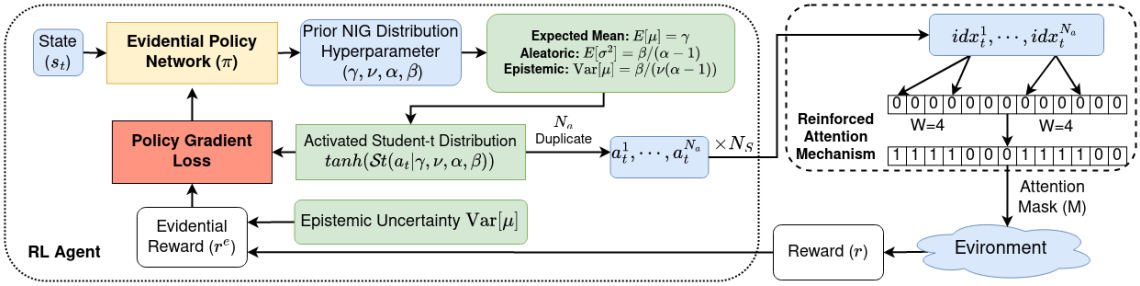


Figure 5: Evidential Reinforced Attentions (ERA)

Similarly, the ERA module finds the optimal policy to maximize long-term expected cumulative evidential reward as:

$$J_\pi(\theta_e) = \sum_{t=1}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t)\sim\pi}(r_\pi^e(\mathbf{s}_t, \mathbf{a}_t)) \tag{14}$$

where T is the total number of time steps in the episode.

For ERA module update, we update the evidential policy network with policy gradient method:

$$\theta_e \longleftarrow \theta_e + \eta_e \nabla_{\theta_e} J_\pi(\theta_e) \tag{15}$$

where $\nabla_{\theta_e} J_\pi(\theta_e)$ is proportion to

$$\mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t)\sim\pi}[r_\pi^e(\mathbf{s}_t, \mathbf{a}_t)\nabla_{\theta_e} \ln \mathrm{St}(\mathbf{a}_t; \pi_{\theta_e}(\cdot|\mathbf{s}_t))]$$

The detailed training process is given in Algorithm 1 of Appendix C. During inference, we input the model with variable length sub-trajectories or full episode trajectories to test the model's capability in classifying the provided trajectory belonging to ASD or TD users. We provide details of the inference process in the experimental section.

5. Experiments

In this section, we conduct both quantitative and qualitative experiments to evaluate the proposed DTS-ERA model on three different games regarding ASD detection, `Maze Painting`, `Word Scanning` and `Coloring`. First, we compare our model with multiple state-of-the-art baselines. The performance comparison shows that our model surpasses all the competitors by a significant margin. We then demonstrate the model's capability in detecting ASD users' unique behaviors using incomplete gameplay trajectories. We use `Maze` task as an example and conduct a detailed ablation study to test the reinforced attention and evidential learning's effectiveness w.r.t. the final classification performance. For qualitative analysis, we show representative SBPs discovered from ASD and TD groups' touch and visual sensory inputs and formally group these SBPs into four major categories. We further conduct a statistical analysis to show that ASD and TD groups of children exhibit a clear distinction on certain SBPs through the Wilcoxon rank sum test. Finally we present a case study to validate the above qualitative analysis results.

Dataset Setup. All data collection experiments are approved by Institutional Review Boards and conducted under informed consent and minor assent. All participants are adolescents (age 11-17 years), and all data are completely deidentified. The `Maze Painting` experiments are carried out with 12 TD children and 12 children with ASD. All participants are seated in front of a screen and played 12 interactive game levels, 6 with 2D mazes and 6 with 3D mazes. The `Word Scanning` and `Coloring` experiments are conducted with 9 children with ASD and 13 TD children. More details about the experiments are in Appendix B.

Comparison baselines. We compare with multiple state-of-the-art baselines, all of which are trained using the same sub-trajectory-based task formulation:

- **Recurrent Classifier** (LSTM) (Hochreiter & Schmidhuber, 1997): We consider a LSTM based classifier as a simple recurrent based baseline that leverages the multimodal sequential information for classification. The model uses two sequence encoders to generate the task representation. Specifically, the output from the final time step is concatenated to obtain the task representation.
- **LSTM-FCN** (Karim et al., 2017): LSTM-FCN explores the augmentation of fully convolutional networks with LSTM RNN sub-modules for time series classification.
- **GRU-FCN** (Elsayed et al., 2018): GRU-FCN replaces LSTM with a gated recurrent unit (GRU) to create a GRU-fully convolutional network hybrid model.
- **1D-ResCNN** (ResCNN) (Zou et al., 2019): ResCNN integrates residual network with convolutional neural network for time series classification.
- **Temporal Convolutional Networks** (TCN): (Bai et al., 2018) TCN extends convolutional neural networks for sequence modeling by introducing dilated, casual convolutional networks for sequential time-series classification.
- **InceptionTime** (Ismail Fawaz et al., 2020): InceptionTime is a scaleable model that utilizes cascade of inception modules for multivariate time series classification.

- **MiniRocket**: (Dempster et al., 2021): MiniRocket transforms input using a fixed set of convolutional kernels and trains a linear classifier using the transformed features.
- **An Explainable Convolutional Neural Network** (XCM) (Fauvel et al., 2021): XCM is a new compact convolutional neural network which extracts information relative to the observed variables and time directly from the input data. Thus, XCM architecture enables a good generalization ability on both large and small datasets.

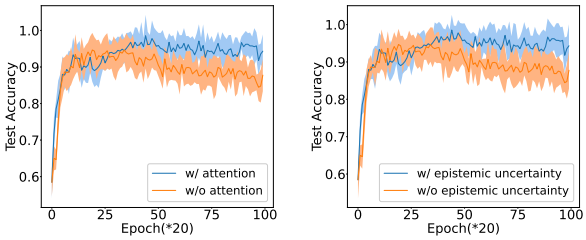
5.1. Quantitative Results

Classification performance. We compare DTS-ERA's predictive accuracy with the baselines on the 2D, 3D and mixed (including both 2D and 3D) game data for `Maze Painting` as well as the other two games in Table 1. As can be seen, DTS-ERA outperforms all baselines by a significant margin. Working across all these real-world datasets further demonstrates our model's generalization capability in human behavior analysis. For component analysis, the basic LSTM model's performance is considerably low as it only looks at the long and short-term temporal structure and fails to capture the representative patterns of autistic users. XCM baseline extracts the complex relationship relative to the observed variables from the input data but fails to leverage the long and short temporal information as LSTM. MiniRocket uses fixed convolutional kernels to obtain the features, and uses a linear classifier for prediction. TCN and InceptionTime models extend the conventional CNN modules for time series data leading to improved classification performance. The DTS baseline shows that even without ERA, our model is stronger than all its competitors by leveraging deep temporal set construction. However, all models lack the novel reinforced attention mechanism to identify the most characteristic behavioral patterns to distinguish ASD from TD. The reinforced attention mechanism enables our model to identify the SBP leading to superior classification performance.

Handling Incomplete Trajectories. Our model can easily be extended to handle situations with partial and incomplete data because our model is a novel extension to the deep sets methods. We test our model's capability in adapting to different length partial trajectory up to the entire episode to match the real cases, where data recording might be partially missing or interrupted. As Figure 7a and 7b show, we use game `Maze Painting` as an example and our model could achieve up to 85% test classification accuracy as the test trajectory going to completion, and already achieves more than 65% accuracy even with only 1/5 completion degree, either in 2D and 3D dataset. Please refer to Appendix E.1 for additional partial trajectory experimental results on the mixed dataset.

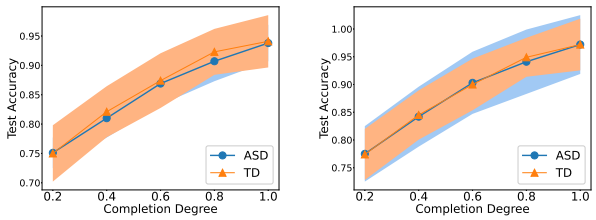
5.2. Ablation Study

We conduct an ablation study with game `Maze Painting` to evaluate two uniquely designed components in our model: epistemic uncertainty and evidential reinforced attentions. As shown in Figure 6a, ERA helps our model identify important SBPs and improve classification accuracy. As shown in Figure 6b, evidential uncertainty-based exploration could help our model identify more diverse patterns from long-term multi-modal data. In addition, Table 2 discloses the newly identified pattern category (in bold font), which supports the effectiveness of the evidential uncertainty-based exploration. In Appendix E.2, we show the effectiveness of different model components like DTS, attentive embedding concatenation or RL mechanism by ablating or analyzing them separately.



(a) Attention (b) Evidential exploration

Figure 6: Average performance w/ and w/o reinforced attention mechanism (a) and evidential exploration (b).



(a) Incomplete data [2D] (b) Incomplete data [3D]

Figure 7: Predictive accuracy for incomplete episodes

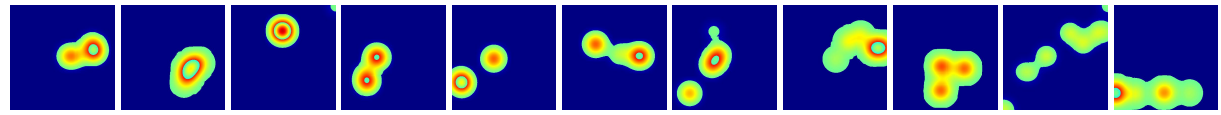
5.3. Qualitative Analysis

Signature Behavioral Patterns. We continue using `Maze Painting` as an illustrative example. Through the reinforced attention mechanism, the RL agent is able to focus on a subset of instances to formulate the state embedding. We find that these selected instances reveal different gaze and touch modality patterns for different user groups, and therefore they can be considered both representative and discriminative. We visualize those patterns using the heatmaps in Figure 8. Each point's location in the heatmap corresponds to a 2-dimensional feature (e.g., the coordinates of the eye gaze position and the ball position on the screen). The brightness corresponds to the gaze lasting time (eye gaze) or pressure on the ball (touch) around that position.

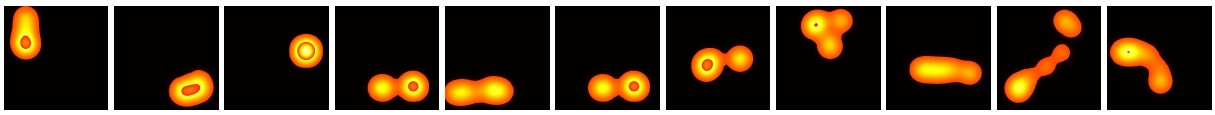
In Figure 9, we show the statistical counts for each sensory

Table 1: Classification Performance Comparison

Dataset	Maze-2D	Maze-3D	Maze-Mixed	Coloring	Word Scanning
LSTM (Hochreiter & Schmidhuber, 1997)	65.0±6.7	63.0±5.5	68.8±4.8	62.0±3.8	62.5±4.1
LSTM-FCN (Karim et al., 2017)	91.5±3.6	93.4±3.5	90.4±3.8	86.0±4.2	89.0±4.1
TCN (Bai et al., 2018)	73.1±4.5	64.0±7.1	60.5±5.7	64.0±4.1	68.0±3.7
GRU-FCN (Elsayed et al., 2018)	92.3±3.2	94.5±4.1	91.8±4.8	88.0±3.9	91.0±3.8
ResCNN (Zou et al., 2019)	85.0±5.2	76.0±4.5	82.0±6.2	85.0±4.3	88.0±4.2
InceptionTime (Ismail Fawaz et al., 2020)	80.1±6.7	72.5±1.6	78.8±4.8	82.0±4.1	77.0±4.2
XCM (Fauvel et al., 2020)	37.5±3.7	44.4±4.3	54.40±3.6	58.0±3.6	52.0±4.2
MiniRocket (Dempster et al., 2021)	70.7±7.1	55.3±3.5	56.3±3.8	65.0±4.8	65.0±5.1
DTS	93.1±3.6	94.6±5.8	92.7±5.6	89.0±4.5	92.5±4.4
DTS-ERA	**94.0±3.8**	**95.3±5.3**	**95.0±5.2**	**91.0±3.9**	**93.5±3.8**



(a) Shift (b) Oval (c) Circle (d) BCS (e) BSS (f) ICS (g) ISS (h) Clutter (i) LS (j) GS (k) LL



(l) Slide (m) Oval (n) Circle (o) BCS (p) BSS (q) ICS (r) ISS (s) Clutter (t) LL (u) GL (v) TA

Figure 8: Heatmap visualizations of representative SBPs in `Maze Painting` across ASD and TD user groups from each discovered gaze [(a)-(k)] and touch [(l)-(v)] pattern category, where all the pattern categories are summarized in Table 2.

Table 2: Categorization of discovered SBPs and Wilcoxon rank sum test for each pattern category

Sensory	Shape	Pattern Category	P-value
Gaze	**Concentrated**	Shift	0.0985
		Oval	0.1042
		Circle	0.0965
	Switching	**Balanced Clutter Switching** (BCS)	0.0675
		Balanced Scatter Switching (BSS)	0.0876
		Imbalanced Clutter Switching (ICS)	0.0752
		Imbalanced Scatter Switching (ISS)	0.0923
	Local	**Clutter**	0.0623
		Local Scan (LS)	0.0518
	Global	**Global Scan** (GS)	0.0432
	Line	Local Line (LL)	0.0447
Touch	**Concentrated**	Slide	0.0802
		Oval	0.0857
		Circle	0.0968
	Switching	Balanced Clutter Switching (BCS)	0.0745
		Balanced Scatter Switching (BSS)	0.0736
		Imbalanced Clutter Switching (ICS)	0.0594
		Imbalanced Scatter Switching (ISS)	0.0654
	Local	Clutter	0.0518
		Turning Around (TA)	0.0485
	Line	Local Line (LL)	0.0469
		Global Line (GL)	0.0423

group's SBPs. From those results, we observe that the ASD group presents more patterns than TD in concentrated gaze patterns (Shift, Oval, Circle) and Switching gaze patterns (BCS, BSS, ICS, ISS), while the TD group reveals more patterns than ASD in local, global and line shape patterns (LS, GS, LL). We further apply a Wilcoxon rank sum test to evaluate if the pattern differences in the two user groups are statistically significant. Here $p < 0.05$ is considered significant (with 95% confidence), while $0.05 \leq p < 0.1$ is considered near significant.

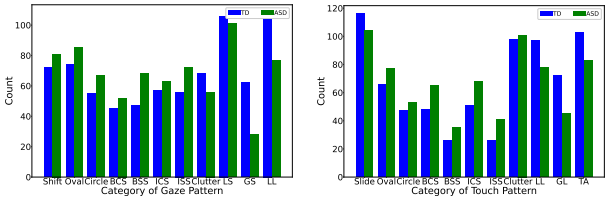
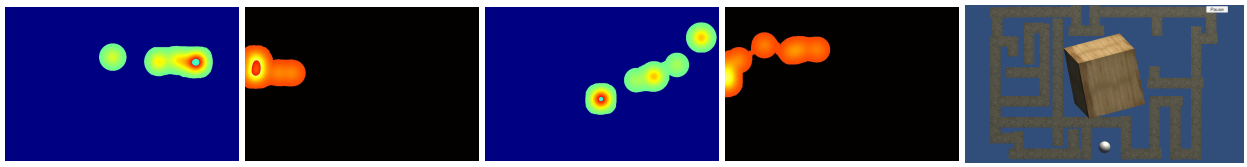


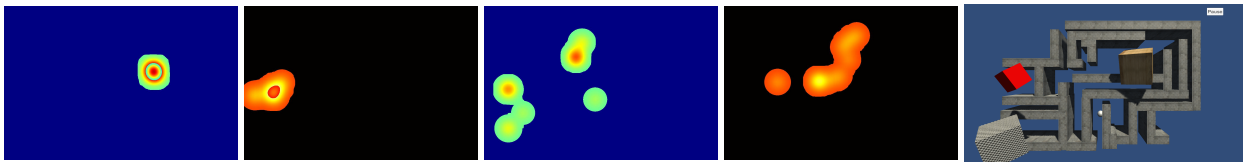
Figure 9: Gaze and touch SBP category statistical analysis Table 2 shows the p-value for each pattern category (the pattern category with a bold font is the sensory pattern detected by the reinforced attention mechanism). From those statistical results, we can claim that the local, global, and line patterns in the ASD and TD groups are significantly or near significantly different. This finding confirms our aforementioned statistical counts. The overall discovery is aligned with previous psychological studies that indicated adolescents with ASD are impaired in responding to stimuli that are subtle or complex (Wieckowski & White, 2020). Case studies on these ASD datasets are provided below that uses the discovered SBPs to better understand autistic children's behaviors.

5.4. Case Study

We conducted a case study in both 2D and 3D `Maze` game environments by collecting autistic and TD users' most frequent gaze and touch SBPs in game levels 6 and 11,



(a) Gaze [ASD]: LL (b) Touch [ASD]: Shift (c) Gaze [TD]: GS (d) Touch [TD]: GL (e) Maze board (2D)



(f) Gaze [ASD]: Circle (g) Touch [ASD]: Slide (h) Gaze [TD]: LS (i) Touch [TD]: TA (j) Maze board (3D)

Figure 10: (a)-(d): gaze and touch SBPs for ASD and TD users in game level 6 (2D-environment); (f)-(j) gaze and touch SBPs for ASD and TD users in game level 11 (3D-environment); (e) and (j) maze boards for 2D and 3D environments

respectively. Level 6 was a 2D game environment with a cube distractor located in the middle; Level 11 was a 3D environment where distractors spread across the board.

The result is shown in Figure 10. In the 2D environment of game level 6 (the first row), both groups' gaze and touch followed a line shape in most cases. The difference was that TD users' line spread longer and more diverse than that of the autistic users, thus forming long-range and curve patterns, such as Global Scan (GS) and Global Line (GL), as compared to autistic users' Local Line (LL) and Shift patterns for gaze and touch. This observation is also consistent with the 3D environment (the second row), where the ASD users' most frequent gaze and touch SBPs were both concentrated patterns. In contrast, the TD users show Local Scan (LS) and Turning Around (TA) patterns for gaze and touch, respectively. We combined 2D and 3D observations and analyzed them jointly and found an interesting phenomenon for autistic users: they tended to look at and move the ball along the math path near the margin of the gaming scene. More importantly, their eye gaze and ball movement in most cases were in the opposite positions, as shown in Figures 10a and 10b as well as 10f and 10g. These were the most frequent gaze and touch patterns of the ASD group in 2D and 3D environments. This finding is aligned with previous child development research that showed children with ASD tend to demonstrate less efficient eye-hand coordination (Crippa et al., 2013). In the 3D environment, autistic users' eye gaze patterns fell more on the cube distractors, shown in Figure 10f. They pushed the ball harder as indicated by brightness in Figure 10g but with little movement. In contrast, the TD users usually scanned more widely beyond the distractors and continued to move the ball along the path near the distractors to complete the painting task, as shown in Figures 10h and 10i. Such observations coincided with our statistical analysis and provided evidences that support previous research about the potential ASD-related impairment in distractor inhibition (Lindor et al., 2019). More qualitative analysis results on the other two tasks (`Word Scanning` and `Coloring`) are included in Appendix E.3.

6. Conclusion

In this paper, we integrate deep temporal sets with evidential reinforced attention to discover subtle and complex multimodal human behavioral patterns. The proposed DTS-ERA model leverages two-tiered sequence encoders (*i.e.,* a temporal encoder and a spatial encoder) to generate DTS embedding that exploits temporal information. For effective model training, DTS-ERA uses the setting of few-shot learning that takes sampled sub-trajectories as the input tasks. By training across different tasks, the model can adapt to user data with different lengths to match real-world scenarios. Additionally, DTS-ERA combines an RL agent to perform evidential reinforced attention that learns an effective policy to select representative embeddings as attention signatures and further boosts the performance. Experimental results on multiple real-world datasets demonstrate the effectiveness of the proposed model. Signature behavioral patterns provide useful and interpretable insights to understand important behaviors that are unique for children with ASD.

Acknowledgement

This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

APA. *Diagnostic and statistical manual of mental disorders (5th ed.)*. American Psychological Association, 2013.

Babrak, L. M., Menetski, J., Rebhan, M., Nisato, G., Zinggeler, M., Brasier, N., Baerenfaller, K., Brenzikofer, T., Baltzer, L., Vogler, C., et al. Traditional and digital biomarkers: two worlds apart? *Digital biomarkers*, 3(2): 92–102, 2019.

Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Bent, B., Wang, K., Grzesiak, E., Jiang, C., Qi, Y., Jiang, Y., Cho, P., Zingler, K., Ogbeide, F. I., Zhao, A., et al. The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mhealth and wearables data. *Journal of clinical and translational science*, 5(1), 2021.

Bozgeyikli, L., Raij, A., Katkoori, S., and Alqasemi, R. A survey on virtual reality for individuals with autism spectrum disorder: design considerations. *IEEE Transactions on Learning Technologies*, 11(2):133–151, 2017.

Cai, Q., Wang, H., Li, Z., and Liu, X. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access*, 7:133583–133599, 2019.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4):1–40, 2021.

Cilia, F., Carette, R., Elbattah, M., Guérin, J.-L., and Dequen, G. Eye-Tracking Dataset to Support the Research on Autism Spectrum Disorder. *Preprint*, 6 2022. doi: 10.6084/m9.figshare.20113592.v1.

Crippa, A., Forti, S., Perego, P., and Molteni, M. Eye-hand coordination in children with high functioning autism and asperger's disorder using a gap-overlap paradigm. *Journal of autism and developmental disorders*, 43:841–850, 2013.

Dempster, A., Schmidt, D. F., and Webb, G. I. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 248–257, New York, 2021. ACM.

Elsayed, N., Maida, A. S., and Bayoumi, M. Deep gated recurrent and convolutional network hybrid model for univariate time series classification. *arXiv preprint arXiv:1812.07683*, 2018.

Fauvel, K., Lin, T., Masson, V., Fromont, É., and Termier, A. XCM: an explainable convolutional neural network for multivariate time series classification. *CoRR*, abs/2009.04796, 2020. URL https://arxiv.org/abs/2009.04796.

Fauvel, K., Lin, T., Masson, V., Fromont, É., and Termier, A. Xcm: An explainable convolutional neural network for multivariate time series classification. *Mathematics*, 9(23):3137, 2021.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. A. Conditional neural processes. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Gondal, M. W., Joshi, S., Rahaman, N., Bauer, S., Wuthrich, M., and Schölkopf, B. Function contrastive learning of transferable meta-representations. In *International Conference on Machine Learning*, pp. 3755–3765. PMLR, 2021.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. E. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

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

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., and Petitjean, F. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

Karim, F., Majumdar, S., Darabi, H., and Chen, S. Lstm fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Koirala, A., Yu, Z., Schiltz, H., Van Hecke, A., Armstrong, B., and Zheng, Z. A preliminary exploration of virtual reality-based visual and touch sensory processing assessment for adolescents with autism spectrum disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:619–628, 2021.

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5): e15708, 2021.

Lee, K. F. A., Gan, W.-S., and Christopoulos, G. Biomarker-informed machine learning model of cognitive fatigue from a heart rate response perspective. *Sensors*, 21(11): 3843, 2021.

Lindor, E., Rinehart, N., and Fielding, J. Distractor inhibition in autism spectrum disorder: Evidence of a selective impairment for individuals with co-occurring motor difficulties. *Journal of Autism and Developmental Disorders*, 49:669–682, 2019.

Maenner, M. J., Shaw, K. A., Bakian, A. V., Bilder, D. A., Durkin, M. S., Esler, A., Furnier, S. M., Hallas, L., Hall-Lande, J., Hudson, A., et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2018. *MMWR Surveillance Summaries*, 70(11):1, 2021.

Mnih, V., Heess, N., Graves, A., and kavukcuoglu, k. Recurrent models of visual attention. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf.

Noel, J.-P., De Niear, M. A., Lazzara, N. S., and Wallace, M. T. Uncoupling between multisensory temporal function and nonverbal turn-taking in autism spectrum disorder. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):973–982, 2017.

Pagnozzi, A. M., Conti, E., Calderoni, S., Fripp, J., and Rose, S. E. A systematic review of structural mri biomarkers in autism spectrum disorder: A machine learning perspective. *International Journal of Developmental Neuroscience*, 71:68–82, 2018.

Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304, 2017. URL http://arxiv.org/abs/1705.04304.

Potamitis, I. and Rigakis, I. Large aperture optoelectronic devices to record and time-stamp insects' wingbeats. *IEEE Sensors Journal*, 16(15):6053–6061, 2016. doi: 10.1109/JSEN.2016.2574762.

Qin, S., Zhu, J., Qin, J., Wang, W., and Zhao, D. Recurrent attentive neural process for sequential data. *arXiv preprint arXiv:1910.09323*, 2019.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Sutton, R. S., Barto, A. G., et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 1999.

Thieme, A., Belgrave, D., and Doherty, G. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53, 2020.

Washington, P., Park, N., Srivastava, P., Voss, C., Kline, A., Varma, M., Tariq, Q., Kalantarian, H., Schwartz, J., Patnaik, R., et al. Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(8): 759–769, 2020.

Wieckowski, A. T. and White, S. W. Attention modification to attenuate facial emotion recognition deficits in children with autism: A pilot study. *Journal of autism and developmental disorders*, 50(1):30–41, 2020.

Ye, J., Xue, Y., Long, L. R., Antani, S. K., Xue, Z., Cheng, K. C., and Huang, X. Synthetic sample selection via reinforcement learning. *CoRR*, abs/2008.11331, 2020.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., and Elsabbagh, M. Global prevalence of autism: a systematic review update. *Autism Research*, 15(5):778–790, 2022.

Zou, X., Wang, Z., Li, Q., and Sheng, W. Integration of residual network and convolutional neural network along with various activation functions and global pooling for time series classification. *Neurocomputing*, 367:39–45, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.08.023. URL https://www.sciencedirect.com/science/article/pii/S0925231219311506.

Appendix

Organization of Appendix

In Appendix A, we summarize the major notations used throughout the paper. In Appendix B, we provide additional details of the datasets and experimental setup. In Appendix C, we show the pseudo code of the training process. Appendix D provides the proofs of the main theoretical results. Additional experiment results are provided in Appendix E. In Appendix F, we discusses the limitations, future work, and social impact of the proposed work. Finally, the link to the source code and processed datasets is provided in Appendix G.

A. Table of Symbols

The major notations and their descriptions are summarized in Table 3.

Table 3: Symbols with Descriptions

Notation	Description
G	Game Experience Collection
$\mathcal{T}_l^u, N_e$	Episode Trajectory for user u in game level l, variable length of episode trajectory
$\mathcal{T}^s, N_s$	Fixed size sub trajectory sampled from episode $\mathcal{T}$ as training input, length of input sub-trajectory $\mathcal{T}^s$
$\mathbf{t}_n, \mathbf{g}_n$	Instance level record in trajectory $\mathcal{T}$ (Instance)
y, y_s	User Category for episode trajectory $\mathcal{T}$ and sub-trajectory $\mathcal{T}^s$
p	padding size due to the kernel size of 1D-Conv layer.
$\{(h^1, ..., h^{N_s})\}$	Multi-modality instance level record in input sub-trajectory $\mathcal{T}^s$
$\{(h_t^1, ..., h_t^{N_s}), (h_s^1, ..., h_s^{N_s})\}$	DTS module generated hidden embedding sets
$\mathbf{d}, \mathbf{d}_{attn}^t$	Average aggregated DTS module generated embedding and RL selected attentive subset embedding
$\mathbf{e}_t$	State Embedding representation
$\mathbf{s}_t$	State space representation in time step t
$\mathbf{a}_t, N_a$	Action space representation in time step t and length of $\mathbf{a}_t$
$r^e(\mathbf{s}_t, \mathbf{a}_t)$	Evidential Reward associated with the state $\mathbf{s}_t$ and action $\mathbf{a}_t$
$r(\mathbf{s}_t, \mathbf{a}_t)$	Tradition Reward calculated by the last time step T's predictive accuracy.
γ_{RL}	Discounting factor used in the cumulative reward computation
T	length of RL time step in one training iteration
P_T	Last time step T's predictive result
W	Attention window length
$idx_t^k, k \in [1, N_a]$	Transform of each relative starting position a_t^k to absolution position index in sub-trajectory $\mathcal{T}^s$.
$\alpha, \beta, \nu, \gamma$	Output of evidential policy network to infer underlying Gaussian distribution.
μ, σ^2	Mean and Variance of the inferred underlying Gaussian distribution for sampling action $\mathbf{a}_t$.
M_t	Binary Attention Mask at time step t
π_{θ_e}, θ_e	evidential policy network and its parameter
$\texttt{NN}, \theta_{NN}$	binary classification network and its parameter
(SE, θ_{se})	State Encoder and its parameter
$\Phi_t(.), \Phi_s(.)$	Temporal and Spatial Sequence Encoders
θ_d	DTS module's joint parameter representation, including the parameter of Sequence and State Encoders.
$\eta_d, \eta_{NN}, \eta_e$	Learning rate of DTS module, classification network and evidential policy network.
λ	Hyper-parameter balancing the exploration and exploitation

B. Additional Details of Datasets and Experimental Setup

During the `Maze Painting` game, each participant finished 6 episodes in 2D scenario and another 6 episodes in 3D scenario. To complete one episode, the participant had to completely paint the maze by operating a painting ball through the Haptic robot. The trajectory of the ball, which reflected the applied touch sensation, was recorded at about 20Hz. In addition, the participant's visual behavior was recorded as gaze position on the screen using a calibrated Tobii Pro x-30 eye-tracker at about 20-30 Hz. During the `Word Scanning` and `Coloring` games, each participant finished 6 episodes in each game by operating the Haptic robot. The trajectory of the cursor (for word scanning) and the pen (for coloring), which reflected the applied touch sensation, was recorded at about 60Hz. The participant's gaze position on the screen was recorded by a calibrated Tobii Pro x-30 eye-tracker at about 20-25 Hz. Each user-episode data consists of a large number of sequential records (one record represents the tracked gaze/touch data at a particular time) and its corresponding length varied across the users and episodes. The tracked records were averaged across 0.1 second sequential intervals to obtain the instances. Averaging over 0.1 second time interval reduced noise in the data and unified the vision and touch sampling rates. Each gaze instance is a 4-dimensional feature vector indexed by time and represents the eye gaze position of each eye on the screen. Each touch instance is a 3-dimensional feature vector indexed by time and represents the ball position on the screen along with the applied pressure. During the data collection process, some users' behavior was not recorded due to hardware/software issues, such as the participants blocking the view of the eye tracker or the robot controller having a glitch. After removing these sections with incomplete data, we have a total of 100 2D user-episode combinations and 125 3D user-episode combinations, leading to a total of 225 user-episode combinations (termed mixed combination). For the other two tasks (`Word Scanning` or `Coloring`), the number of user-episode combinations are both 132. We consider 60% of the available data (user episodes) of each sub-group (2D, 3D, mixed or word scanning, coloring) for training, train the model for 2000 epochs, and evaluate the model on the remaining data in that sub-group. We repeat those experiments across 5 random train-test splits and average the test set accuracy for quantitative evaluations.

C. Training Algorithm

In this section, we first derive the posterior predictive distribution of evidential policy network (i.e. Eq. 7) and then provide the Algorithm 1 for detailed training process.

The evidential policy network is set up in such a way that the target, *i.e.,* action $\mathbf{a}_t$, is drawn i.i.d. from a Gaussian distribution with unknown mean and variance (μ, σ^2). Model evidence can be introduced by further placing a prior distribution on (μ, σ^2). To ensure conjugacy, we choose a Gaussian prior on the unknown mean and an Inverse-Gamma prior on the unknown variance:

$$p(\mathbf{a}_t|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) \tag{16}$$

$$p(\mu|\gamma, \sigma^2\nu^{-1}) = \mathcal{N}(\gamma, \sigma^2\nu^{-1}) \tag{17}$$

$$p(\sigma^2|\alpha, \beta) = \text{Inv-Gamma}(\alpha, \beta) \tag{18}$$

The joint posterior of (μ, σ^2) can be formulated as a Normal Inverse-Gamma (NIG) distribution:

$$p(\mu, \sigma^2|\gamma, \nu, \alpha, \beta) = \frac{\beta^\alpha\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left\{-\frac{2\beta + \nu(\gamma-\nu)^2}{2\sigma^2}\right\} \tag{19}$$

We can interpret the NIG distribution as a higher-order, evidential distribution, where the parameters of this distribution can be interpreted as the (prior) pseudo-observations of an action. Our policy network directly predicts the NIG parameters, which enables us to evaluate evidence for each action and measure epistemic uncertainty.

By marginalizing over (μ, σ^2), we arrive at the following predictive student-t distribution, which is used to sample actions:

$$
\begin{aligned}
p(\mathbf{a}_t|\gamma,\nu,\alpha,\beta) &= \int_{\sigma^2=0}^{\infty}\int_{\mu=-\infty}^{\infty} p(\mathbf{a}_t|\mu,\sigma^2)p(\mu,\sigma^2|\gamma,\nu,\alpha,\beta)\mathrm{d}\mu\mathrm{d}\sigma^2 \\
&= \int_{\sigma^2=0}^{\infty}\int_{\mu=-\infty}^{\infty}\left[\sqrt{\frac{1}{2\pi\sigma^2}}\exp\{-\frac{(a_t-\mu)^2}{2\sigma^2}\}\right]\left[\frac{\beta^\alpha\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\{-\frac{2\beta+\nu(\gamma-\mu)^2}{2\sigma^2}\}\right]\mathrm{d}\mu\mathrm{d}\sigma^2 \\
&= \int_{\sigma^2=0}^{\infty}\frac{\beta^\alpha\sigma^{-3-2\alpha}}{\sqrt{2\pi}\sqrt{1+\frac{1}{\nu}}\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\{-\frac{2\beta+\frac{\nu(a_t-\mu)^2}{1+\nu}}{2\sigma^2}\}\mathrm{d}\sigma^2 \\
&= \int_{\sigma=0}^{\infty}\frac{\beta^\alpha\sigma^{-3-2\alpha}}{\sqrt{2\pi}\sqrt{1+\frac{1}{\nu}}\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\{-\frac{2\beta+\frac{\nu(a_t-\mu)^2}{1+\nu}}{2\sigma^2}\}2\sigma\mathrm{d}\sigma \\
&= \frac{\Gamma(1/2+\alpha)}{\Gamma(\alpha)}\sqrt{\frac{\nu}{\pi}}(2\beta(1+\nu))^\alpha(\nu(a_t-\gamma)^2+2\beta(1+\nu))^{-(\frac{1}{2}+\alpha)} \\
&= \mathrm{St}\left(\mathbf{a}_t;\gamma,\beta(1+\nu)/(\nu\alpha),2\alpha\right)
\end{aligned}
\tag{20}
$$

Algorithm 1 Model Training

Require: Episode Trajectory $\mathcal{T}$ of length N_e sampled from Game Experience Collection $G = \{T_u^l\}, u \in U, l \in L$, total iteration number K, total time step T

Require: Hyperparameters: η_d, η_{NN}, and η_e (learning rates), γ_{RL} (RL discount factor), evidential policy network parameters (θ_e), Deep Temporal Sets module parameters (θ_d), and classification network parameters (θ_{NN})

while iteration k $\leq$ K **do**

Randomly sample Episode Trajectory $\mathcal{T}$ from Game Experience Collection and sample sub-trajectory $\mathcal{T}^s$ of N_s instances from $\mathcal{T}$

for time step t $\in$ [1,T] **do**

Input sub trajectory $\mathcal{T}^s$ into DTS module to generate gaze and touch representation as (3).

Concatenate DTS module generated embeddings and RL-agent selected attentive regions as in (5).

Input e_t into state encoder as equation (6) to generate current state $\mathbf{s}_t$

Input state $\mathbf{s}_t$ into evidnetial policy network to get action $\mathbf{a}_t$, and calculate evidential reward $r^e(\mathbf{s}_t, \mathbf{a}_t)$ using equation (9) in the last time step T

Pass $\mathbf{s}_t$ to classification network to get final prediction p_T and update the network as (13)

end for

Update evidentail policy network as (15) and DTS module as (12) respectively

end while

D. Proofs of Theoretical Results

In this section, we provide proofs of all lemmas and the theorem.

D.1. Proof of Lemma 1

Proof. Given the evidential reward defined as $r^e(\mathbf{s}_t) = r_\pi(\mathbf{s}_t) + \lambda\texttt{epistemic}_\pi(.|\mathbf{s}_t))$ the update rule for epistemic value can be written as:

$$
V^e(\mathbf{s}_t) = \mathbb{E}_\pi\sum_{t'=t}^{\infty}\gamma^{t'}r^e(\mathbf{s}_{t'}) = r^e(\mathbf{s}_t) + \gamma\mathbb{E}_{\mathbf{s}_{t+1}}[V^e(\mathbf{s}_{t+1})] \tag{21}
$$

Following the convergence rule (Sutton et al., 1999) with finite action space, it is guaranteed that the value will converge to the epistemic value of policy π. □

D.2. Proof of Lemma 2

Proof. The policy can be updated towards the new value function. Consider the updated policy π_{new} as the optimizer of the maximization problem.

$$\pi_{new} = \arg\max_{\pi'} J_\pi(\phi) = \arg\max_{\pi'} \mathbb{E}_{\mathbf{s}_t}[V^e_{\pi'}(\mathbf{s}_t)] \tag{22}$$

Denote the old policy as π_{old}. Using the update rule specified in Eq (15) with a sufficiently small step size, we get an updated policy π_{new} that satisfies

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{new}}[V^e_{\pi_{old}}(\mathbf{s}_t)] \geq \mathbb{E}_{\mathbf{a}_t\sim\pi_{old}}[V^e_{\pi_{old}}(\mathbf{s}_t) \tag{23}$$

Given Eq (23), we have the following inequality

$$\begin{aligned} V^e_{\pi_{old}}(\mathbf{s}_t) \leq & r^e(\mathbf{s}_t) + \gamma\mathbb{E}_{\mathbf{s}_{t+1},\mathbf{a}_{t+1}\sim\pi_{new}}[V^e_{\pi_{old}}(\mathbf{s}_{t+1})] \\ \leq & r^e(\mathbf{s}_t) + \gamma\mathbb{E}_{\mathbf{s}_{t+1},\mathbf{a}_{t+1}\sim\pi_{new}}[r^e(\mathbf{s}_{t+1})] \\ & + \mathbb{E}_{\mathbf{s}_{t+2},\mathbf{a}_{t+2}\sim\pi_{new}}[V^e_{\pi_{old}}(\mathbf{s}_{t+2})] \\ & \ldots \\ = & V^e_{\pi_{new}}(\mathbf{s}_t) \end{aligned} \tag{24}$$

where $r^e(\mathbf{s}_t)$ is a evidential reward in step t. Therefore, we show that the new policy π_{new} ensures $V^e_{\pi_{new}}(\mathbf{s}_t) \geq V^e_{\pi_{old}}(\mathbf{s}_t)$ for all $\mathbf{s}_t$. □

D.3. Proof of Theorem 3

Proof. Let π_i denote the policy at iteration i. We already show that the sequence $V^e_{\pi_i}(\mathbf{s}_t)$ is monotonically increasing. Since $V^e_\pi(\mathbf{s}_t)$ is bounded above, the sequence converges to some π^*. At convergence, it must be the case that $J_{\pi^*}(\pi^*(.|\mathbf{s}_t)) \leq J_{\pi^*}(\pi(.|\mathbf{s}_t))$ for $\pi \neq \pi^*$. Based on Lemma 2, we have $V^e_{\pi^*}(\mathbf{s}_t) > V^e_\pi(\mathbf{s}_t)$ for all $\mathbf{s}$. In other words, the evidence value of any other policy π is lower than that of the converged policy π^*. Therefore, it guarantees convergence to an optimal policy π^* such that:

$$V^e_{\pi^*}(\mathbf{s}_t) \geq V^e_\pi(\mathbf{s}_t) \tag{25}$$

□

E. Additional Experiment Results

In this section, we present additional empirical evaluation, including more experimental results, ablation study, evaluation on additional datasets, and qualitative results.

E.1. More Experimental Results

E.1.1. Comparison on additional datasets

To further demonstrate the generalization ability of the model, we conduct experiments on two additional public datasets, including one dataset that records the eye movements from children with and without ASD (Cilia et al., 2022) and another large-scale time-series dataset that contains wingbeat recordings of six mosquito species (Potamitis & Rigakis, 2016). The accuracy results are provided in Table 4, which also include two best performing baselines, LSTM-FCN and GRU-FCN. The result confirms that the model achieves consistently better generalization performance than the competitive baselines.

E.1.2. Statistical test on comparison results

In this set of experiments, we conduct a statistical test to demonstrate that the proposed DTS-ERA model performs better than the other baselines with sufficient statistical evidence. In particular, we train our model and two strongest baselines *GRU-FCN* and *LSTM-FCN* 5 times using randomly sampled 60% training data and test the performance on the remaining 40% data in each sub-group. Based on the results, we conduct the Wilcoxon test and Table 5 reports the result. As can be seen, all the p-values are no greater than 0.1 with 50% of the entries no more than 0.05. This confirms that the proposed model achieves better prediction performance than the two most competitive baselines with sufficient statistical significance.

Table 4: Model Comparison on WingBeats and Eye-tracking Datasets

Model	Eye-tracking	WingBeats
LSTM-FCN	0.69	0.92
GRU-FCN	0.70	0.935
DTS	0.735	0.96
DTS-ERA	**0.758**	**0.975**

Table 5: p- values from the Wilcoxon test over two strongest baselines *GRU-FCN* and *LSTM-FCN*

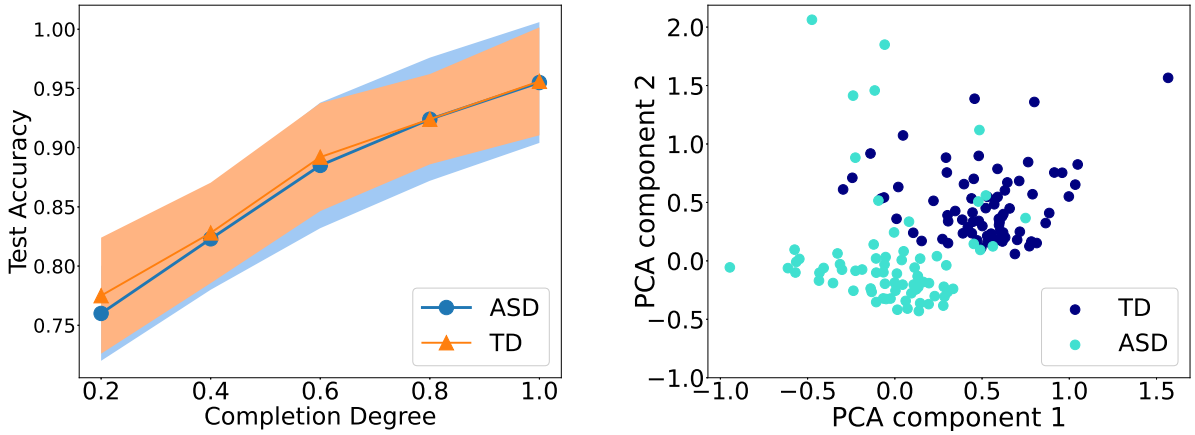
Model	Maze-2D	Maze-3D	Maze-Mixed	Coloring	Word Scanning
LSTM-FCN	0.06	0.07	0.04	0.03	0.05
GRU-FCN	0.07	0.10	0.05	0.05	0.07

E.1.3. Additional Result on Incomplete Trajectories

Figure 11a shows DTS-ERA's performance by taking 20%, 40%, 60%, 80% and whole trajectory on the mixed dataset (*i.e.,* 2D+3D). This result is consistent with the results shown in the main paper, which further demonstrates our model's good detection performance using only partial trajectories.

E.1.4. Visualization of DTS-ERA Embedding

To provide additional insights on the outstanding detection performance of the proposed model, we project the learning DTS-ERA selected attentive subset embedding from the sampled sub-trajectory $\mathcal{T}^s$ of both groups in Maze-Mixed dataset into a low-dimensional space and visualize the testing trajectories in Figure 11b. We leverage PCA to further reduce the dimensionality of attentive subset embeddings and choose two principle components to visualize the trajectories (each is mapped to a point in the 2d space). As can be seen, the TD and ASD groups become easily separated even in such a low-dimensional space.



(a) Incomplete Data [Mixed]

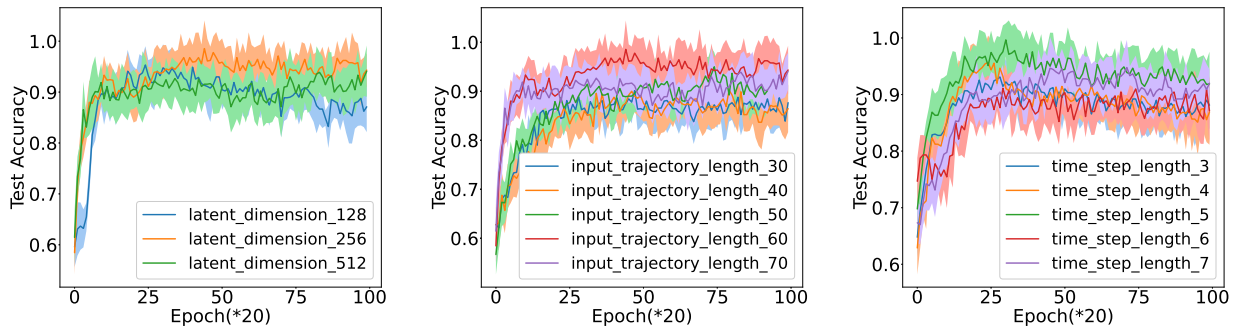
(b) Visualizing the DTS-ERA embedding

Figure 11: Performance of incomplete input data trajectory (a) and Visualization of the DTS-ERA embedding (b) in Maze-Mixed dataset.

E.2. Additional Ablation Study

E.2.1. Impact of Hyper-parameters

Figure 12 presents the impact of different hyper-parameters: dimensionality of DTS embedding, length of the input sub-trajectory N_s, and length of RL time step T in one training iteration. Specifically, when embedding's latent dimensionality is set to 256, input sub-trajectory length is set to 60, and RL time step is set to 5, our model achieves the best performance.



(a) DTS embedding dimensionality (b) Input sub-trajectory length (c) RL time step length

Figure 12: Impact of hyper-parameters

E.2.2. Impact of Deep Temporal Sets

Before generating the aggregated state embedding, two important steps occur that demonstrate the key advantages of DTS. First, a deep set is constructed where each entry in the set is created to record the most important patterns that occur up to that time step. In this way, DTS can effectively capture all the important patterns including those that occur much earlier in a sequence. For LSTM or GRU-inspired baselines, they mainly leverage the final output in last unit as feature embedding. Although it encodes historical information to some extent, it may still forget important information especially when the time sequence becomes long. By further integrating with the spatial encoder, DTS can more effectively capture important information from the space dimension. Second, by maintaining the individual entries in the deep set, DTS allows the RL agent to attend to the most important entries in the set. By performing time alignment with the original sequence, we can easily recover the corresponding user behaviors that allow us to interpret the meaning of the discovered behavioral patterns. Interpretability is a critical requirement for human behavioral/mental health study. Using LSTM or GRU-inspired models does not offer such a good interpretability. To demonstrate the advantage of DTS, we compare LSTM and GRU baselines with our DTS module and the full DTS-ERA model in the Table 6 below. The results show that when using the DTS module alone, it already achieves a clear advantage. Please note that LSTM, GRU and our models are in the same best configuration, where the sub-trajectory length is 50 and dimensionality is 256.

Table 6: Comparison among full DTS-ERA, DTS module (only) and two other baselines (LSTM and GRU).

Model	Maze-2D	Maze-3D	Maze-Mixed	Coloring	Word Scanning
LSTM	65.0±6.7	63.0±5.5	68.8±4.8	62.0±3.8	62.5±4.1
GRU	66.1±6.3	64.2±5.0	70.2±4.1	63.8±3.2	64.4±3.3
DTS	**93.1±3.6**	**94.6±5.8**	**92.7±5.6**	89.0±4.5	92.5±4.4
DTS-ERA	**94.0±3.8**	**95.3±5.3**	**95.0±5.2**	**91.0±3.9**	**93.5±3.8**

E.2.3. Impact of RL-based Attention Mechanism

We conduct an additional experiment using a standard (not using RL) attention mechanism adopted by existing models (*e.g.,* Attentive Neural Process or ANP) (Kim et al., 2019). We combine such an attention (referred to as ANP) with DTS. Table 7 reports the comparison results. It shows that the proposed evidential reinforced attention (ERA) achieves a better prediction performance on all the datasets. In addition, the proposed RL-based attention also allows us to locate important signature

behavioral patterns that makes the model's output much more interpretable. In contrast, the standard attention performs a weighted aggregation of all the entries in the deep set, which does not offer a good interpretability.

Table 7: Comparison between DTS-ANP and DTS-ERA that leverages RL-based attention mechanism

Model	Maze-2D	Maze-3D	Maze-Mixed	Coloring	Word Scanning
DTS-ANP	92.8±3.4	94.5±5.9	91.8±4.7	88.0±4.4	92.2±4.2
DTS-ERA	**94.0±3.8**	**95.3±5.3**	**95.0±5.2**	**91.0±3.9**	**93.5±3.8**

We also conduct experiments by only leveraging the patterns chosen by ERA to make the final prediction without concatenating other patterns. The result from Table 8 shows a performance gap compared to the complete model due to its inability to access other supporting behavioral information as the necessary context.

Table 8: Comparison between attentive embedding (only) and full DTS-ERA model.

Model	Maze-2D	Maze-3D	Maze-Mixed	Coloring	Word Scanning
DTS-ERA (attention only)	92.8±2.4	94.4±3.8	92.4±3.8	88.5±2.5	92.1±2.5
DTS-ERA	**94.0±3.8**	**95.3±5.3**	**95.0±5.2**	**91.0±3.9**	**93.5±3.8**

E.3. More Qualitative Analysis

Given the simpler gaming operations in `Word Scanning` and `Coloring`, the SBPs found on those tasks are rather simple. Both ASD and TD groups demonstrated a certain level of consistency between touch and gaze modalities, while autistic children's gaze and touch locations seem more separated than the TD group. TD users' gaze and touch patterns were more likely to show (global) line, curve or (local) clutter patterns, while the ASD group tended to show more localized patterns, such as **concentrated** and **switching**. We choose one most frequent gaze or touch SBPs in `Word Scanning` and `Coloring` tasks respectively to demonstrate this finding, shown as Figure 13.

(a) ASD gaze and touch SBPs on word scanning game (b) TD gaze and touch SBPs on word scanning game (c) ASD gaze and touch SBPs on coloring game (d) TD gaze and touch SBPs on coloring game

Figure 13: SBPs visualization for word scanning and coloring games

F. Limitations, Future Work, and Social Impact

While the proposed new DTS-ERA model presents a general method for complex behavioral pattern discovery, our evaluation focuses on ASD-related behavior detection using datasets collected from child-computer interaction. Nevertheless, the model evaluation methodology can be adapted to other datasets and baseline algorithms in a broader context. It is worth to note that the current datasets have missing data, which is a common issue in real-world scenarios due to human and device limitations. Thanks to the few-shot training approaches, our model still achieves robust prediction performance. The current discoveries emphasize potential applications of the model for ASD assessment. This is a starting point for future extensions of our framework: leverage machine learning technology to help improve human behavioral pattern discovery in real-world applications, such as mental/behavioral health care. As a next step, we will utilize more diverse and longer-time human behavior data within and beyond the ASD field to further explore real-world challenges and adapt the proposed model to more complex settings. The proposed DTS-ERA model is generally applicable to many sequential decision-making real-world problems such as healthcare, robotics, and gaming domains. For instance, the model could potentially be used for preliminary screening before expert (such as healthcare professionals) validation, which may speed up disease diagnosis and provide patients with prompt and effective treatment. Moreover, our model could also be used for digital behavior biomarker

detection and digital behavioral phenotyping to better understand the behavioral/sensory patterns of individuals with mental and/or behavioral issues. These will further advance diagnosis and treatment.

G. Source Code

The source code and processed datasets can be accessed here: https://github.com/wdr123/DTS_ERA