MODELING NEURAL ACTIVITY WITH TRANSFORMERS TO PREDICT IMPULSIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Impulsivity is a key behavioral concern associated with numerous mental health disorders and attention-deficit/hyperactivity disorder (ADHD) in particular. Assessment of impulsivity traditionally relies on rating scales, particularly in the clinical setting. These measures have known limitations due to their subjective nature, including potential to be affected by recency effects, cultural bias, contextual factors, and their inability to assess underlying cognitive processes. As a consequence, there has been a long-standing effort to identify the biological basis of impulsivity, using neuroimaging techniques such as functional MRI (fMRI). We propose a machine learning approach that integrates behavioral measures of impulsivity and reward sensitivity with task-based fMRI data to identify patterns of brain activation associated with impulsivity in a group highly enriched with impulsivity and ADHD. Using a Win/Loss reward-processing task, we extracted regression coefficients (beta values) from a Generalized Linear Model applied to fMRI time series within Regions of Interest (ROIs) selected based on prior meta-analysis. The beta values, combined with spatiotemporal embeddings, were input to a transformer encoder to learn latent representations associated with high versus low impulsivity. We trained and validated the model, and further examined its internal reasoning using attention maps. The model achieves effective classification accuracy in distinguishing low versus high impulsivity across different subject groups, including adolescents and young adults—ages typically associated with greater impulsiveness and risk-taking. Furthermore, the attention maps show correspondence to the current understanding of the neural basis of reward processing and impulsivity in key ROIs. This work demonstrates the feasibility of applying transformers to fMRI-based tasks and is a promising tool to identifying patterns of brain activity associated with complex behavioral constructs with clinical importance.

1 INTRODUCTION

Impulsivity, described as acting without thinking, poses public health risks, as it is associated with greater suicidality, higher substance abuse, lower educational and work achievements, poorer physical health, and higher accident rates. In attention-deficit/hyperactivity disorder (ADHD), impulsivity is considered a core characteristic American Psychiatric Association (2022). Individuals with ADHD frequently struggle with immediate reactions and may find delaying gratification particularly challenging. Beyond ADHD, impulsivity has also been implicated in a range of psychiatric conditions, including substance use disorders, mood disorders, and conduct problems, highlighting its broad relevance for clinical research and intervention Bakhshani (2014); Barkley (2015); Faraone et al. (2021); O'Grady & Hinshaw (2021); Hinshaw et al. (2012).

Given its clinical importance, accurate assessment of impulsivity is critical. Yet psychiatric diagnoses, including those for ADHD, often rely on subjective clinical evaluations. Typically, this is in the form of rating scales completed by parents, teachers or self-report, for adults. While rating scale reports provide valuable information, they can also introduce variability, potential bias based on one's individual frame of reference, and are highly affected by recency effects. As a result, individuals may be over- or under-diagnosed, leading to unnecessary interventions or missed opportunities for timely treatment, both of which can carry significant psychological, medical, and social consequences Pierre (2013). Furthermore, they are limited in how well they can explain the neural

underpinnings associated with impulsivity. However, laboratory measures of impulsivity in clinical populations, such as ADHD do not correlate well with rating scales Barkley (2019). Thus, there is a pressing need for objective, reliable, and biologically grounded markers that can supplement traditional assessments and improve diagnostic precision.

To deepen our understanding of psychiatric disorders and improve diagnostic accuracy, research has focused on developing biologically-grounded tools. One approach integrates behavioral assessments with neuroimaging, particularly functional magnetic resonance imaging (fMRI). As a non-invasive method for measuring brain activity, fMRI offers unique insights into the neural mechanisms underlying psychiatric symptoms. In ADHD, fMRI research has identified abnormalities in both reward processing circuits and attention/executive control networks, particularly during tasks involving reward anticipation and attentional control Oldham et al. (2018); Fassbender & Schweitzer (2006); Faraone et al. (2021). These findings suggest that fMRI can provide mechanistic evidence of the neural underpinnings of impulsivity, complementing behavioral measures with direct indicators of brain function. By combining neuroimaging data with computational models, we can move toward more precise, personalized, and mechanistically informed diagnoses and interventions.

Despite the promise of fMRI for uncovering neural mechanisms underlying psychiatric symptoms, interpreting the data remains challenging. The Blood Oxygen Level Dependent (BOLD) signal is noisy, high-dimensional, and spatio-temporally structured, requiring models that can capture subtle patterns across both space and time. Traditional statistical methods Bzdok (2017) often fall short in handling this complexity, especially when the goal is to predict individual differences or identify clinically meaningful subgroups. These limitations have spurred the adoption of machine learning (ML) approaches, which can leverage the full richness of the data by learning directly from raw or minimally processed signals Zhu et al. (2023). In particular, modern deep learning models, such as transformers, are well suited to discover distributed and context-dependent patterns that might be invisible to conventional analyses.

The goal of this study is to integrate clinical concerns about diagnostic reliability with advanced computational methods capable of uncovering the neural mechanisms of impulsivity. Towards this broader goal, in this paper, we combine objective brain imaging data, behaviorally relevant tasks, and transformer-based deep learning models to address two research questions: 1) can we achieve reasonable accuracy in classifying subjects into high and low impulsivity, and 2) can we correlate the context information encoded in the attention maps with corresponding neuroscientific understanding of neural activation to interpret the model's output? We answer both questions in the affirmative. In doing so, we take a step toward more robust tools for understanding impulsive behavior and improving diagnostic precision in psychiatry.

2 Related Work

Task-based fMRI analysis requires structuring high-dimensional brain responses to capture both spatial and temporal patterns relevant to cognitive tasks. A common approach is to reduce voxel-level complexity by aggregating data within regions of interest (ROIs) using brain atlases. This facilitates the study of temporal dynamics while improving interpretability Poldrack (2007); Craddock et al. (2012). Alternatively, data can be summarized across time using beta estimates from regression analyses, which quantify response magnitudes for specific task conditions Monti (2011). Although these approaches are effective for summarizing responses at the condition level, they may smooth over subtle spatial patterns within regions and temporal dynamics within conditions.

To address these limitations, more flexible modeling frameworks have been explored. Deep learning methods, for example, combine convolutional neural networks (CNNs) for spatial feature extraction with recurrent architectures such as RNNs or LSTMs for temporal modeling Huang et al. (2021). Although these approaches improved predictive accuracy, they often struggled with vanishing or exploding gradients, high memory demands, and training instability Bengio et al. (1994); Pascanu et al. (2013).

More recently, attention-based transformer architectures have shown promise in fMRI analysis Deng et al. (2022). Transformers excel at modeling long-range dependencies and context-aware relationships without recurrence Dosovitskiy et al. (2020), achieving state-of-the-art results in psychiatric classification tasks Dai et al. (2024); Cong et al. (2024). However, most transformer-based studies

emphasize performance metrics (e.g., accuracy, AUC-ROC) while offering limited interpretability regarding the neural mechanisms driving their predictions Rudin (2019); Munroe et al. (2024).

Our study addresses this gap by applying a transformer-based model that integrates spatial, temporal, and regional embeddings (Figure 2). This approach not only leverages transformers' ability to capture complex spatio-temporal dependencies but also facilitates interpretability using attention maps, allowing us to probe the neural activations underlying impulsivity in adolescents and young adults with and without ADHD.

3 METHODOLOGY

3.1 DATA METHODOLOGY

We utilized data from the project Mapping Impulsivity's Neurodevelopmental Trajectories (MINT; R01MH091068), which integrates multimodal neuroimaging and clinical assessments to investigate the neurodevelopmental pathways of impulsivity in adolescents and young adults. The MINT project recruited adolescents and young adults as impulsivity and risk-taking are heightened during this developmental period Steinberg et al. (2018); Chase et al. (2017). See references for diagnostic procedures Mukherjee et al. (2022); Kahle et al. (2021); Mukherjee et al. (2021); Elliott et al. (2022); Elahi et al. (2024); Komijani et al. (2025).

To investigate reward processing in ADHD, a central consideration was selecting paradigms that reliably elicit reward-related brain responses. Impulsivity is typically studied within the context of reward tasks because impulsiveness is potentiated in the context of actions related to reward and thus reward tasks can directly probe a behavioral manifestation of impulsiveness. To address this, we employ a Win/Loss task D'Ardenne et al. (2008) designed to probe reward processing with minimal reliance on cognitive control. This task-based approach is particularly well-suited for ADHD research, where deficits in reward responsiveness are thought to be more fundamental than those in executive function.

In analyzing fMRI data, we adopt a top-down, theory-driven approach focused on predefined ROIs. This strategy, guided by existing anatomical and functional knowledge, limits the number of statistical comparisons, thereby reducing the risk of false positives and enhancing the reliability of findings Lieberman et al. (2009); Sauvayre (2023). Importantly, it also mitigates the risk of post-hoc selection bias, a concern raised during the "voodoo correlations" debate in neuroimaging Vul et al. (2009). Subsequent clarifications have emphasized that well-defined, hypothesis-driven ROI analyses can yield valid and interpretable results when executed rigorously Poldrack (2007).

To define our ROIs, we selected eight regions based on significant activation peaks reported in a reward outcome meta-analysis by Oldham et al. Oldham et al. (2018). These included the ventral striatum (VStr; right: 12, 10, -10; left: -14, 8, -28), the Orbitofrontal/Ventromedial Prefrontal Cortex (OFC/vmPFC; right: 2, 44, -10; left: -22, 42, -6; left anterior: -26, 52, -14), the Amygdala (AMYG; right: 22, -22, -14; left: -18, 0, -16), and the Posterior Cingulate Cortex (PCC; right: 2, -36, 36) (see Table B1 in the Appendix for full ROI mappings). These ROIs were selected for their robust and well-documented association with reward outcome processing and their relevance to impulsivity. Coordinates are given in millimeters (mm) relative to the Montreal Neurological Institute template.

3.1.1 WIN/LOSS PARADIGM

The imaging data were collected while participants completed the Win/Loss paradigm, a task designed to assess reward-related neural processes D'Ardenne et al. (2008), which are known to be altered in ADHD and minimally dependent on cognitive control. Imaging acquisition parameters and preprocessing details are described in Appendix A.

The Win/Loss paradigm (Figure 1) is an event-related design consisting of four runs, each beginning with a 4s fixation period followed by 15 trials. In each trial, participants are presented with a single-digit number (0–9) for 2s, during which they must make a forced-choice button press to guess whether a hidden number is higher or lower than the presented digit. This phase serves as the *reward cue period*. A 2s response period follows, and the trial concludes with a 2s outcome period during which participants receive feedback: wins are indicated by a smiling emoji-style face and a $1

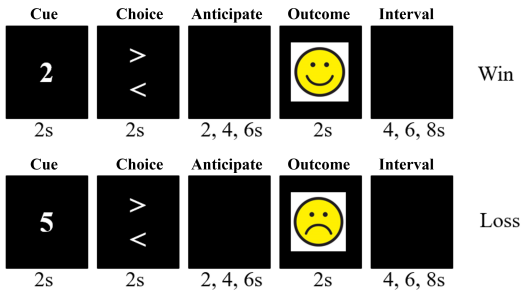

Figure 1: Win/Loss Paradigm: 1) Cue: presentation of a single digit number 2) Choice: the participant makes a guess whether the next number is higher or lower 3) Anticipate: participants anticipates reward or loss 4) Outcome: the smiley or sad face is shown to the participant 5) An inter-trial interval.

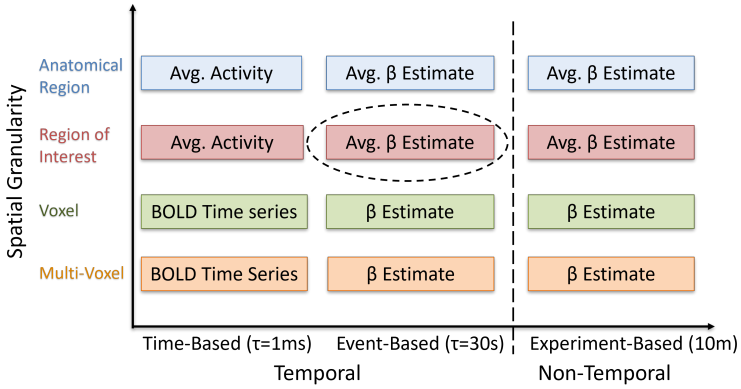

Figure 2: Summary of spatial and temporal resolution levels used for modeling BOLD signal. Spatial units range from anatomically defined brain regions to individual voxels and multi-voxel patterns. Temporal aggregation spans raw BOLD time series, event-level beta values, and experiment-averaged beta values.

reward, while losses are shown by a frowning face and a $0.50 deduction. Inter-trial intervals (ITIs) were jittered at 4, 6, or 8s. Participants received a percentage of their total earnings as compensation.

We calculated the beta values per ROI for task events. There are two primary task events: CUE and OUTCOME. The CUE corresponds to the presentation of the first number, when participants make their guess and form expectations about the likelihood of being correct. The OUTCOME occurs when the result of the guess is realized and a reward or penalty is received. These beta values represent the estimated amplitude of the BOLD response associated with each event and were obtained by fitting a general linear model (GLM) Monti (2011) to the fMRI time series data using Analysis of Functional NeuroImages (AFNI) tool Cox (1996). For each ROI, the mean beta value was extracted using subject-specific masks, providing a measure of neural activation during each task phase.

3.1.2 Clinical Data

To calculate an impulsivity score, we combined multiple measures. This included rating scale measures of impulsivity, based on the Diagnostic and Statistical Manual (DSM)-based hyperactivity and impulsivity criteria American Psychiatric Association (2022) such as the parent Conners' Parent Rating Scale-3 (CPRS; for youth) or observer (could be parent, close friend, spouse, etc.) or self (whichever was higher) from the Conners Adult ADHD Rating Scale (CAARS). We also included hedonism and future-oriented subscales from the Zimbardo Time Perspective Inventory (ZTPI) Zimbardo & Boyd (1999), motor impulsivity and self-control subscales from the Barratt Impulsiveness

Scale (BIS) Patton et al. (1995), and performance-based indices from the Balloon Analogue Risk Task (BART) Lejuez et al. (2002) and delay discounting (DD) task Kirby et al. (1999) measures. We applied factor analysis Ghojogh et al. (2021) to identify latent constructs across these impulsivity-related traits. The first factor, which explained the largest portion of shared variance, loaded strongly on DSM CPRS/CAARS, motor impulsivity, self-control, and hedonism (with a negative loading on future orientation), and was interpreted as a general impulsivity dimension. Factor scores from this latent factor were used as a composite impulsivity score in subsequent analyses.

The results in this paper are based on fMRI data and clinical measures from 182 participants (86 female), aged 12 to 30 years (mean=18.06, SD=4.41). The sample included individuals with ADHD, Combined Presentation, and typically developing individuals. In addition, we divided the participants into two groups of 99 adolescents (37 female) aged 12 to 18 (mean=14.63, SD=1.56) and 83 young adults (49 female) aged 18 to 30 (mean=22.19, SD=2.93). Splitting based on age was done due to known functional and anatomical differences that occur during maturation. Table 1 shows the impulsivity label distribution for all subjects derived from factor analysis.

Table 1: Group distribution of participants classified as low versus high impulsivity based on factor analysis.

Subject Group	Low	High	Total	Class Ratio (Low:High)
Adolescents	58	41	99	1:0.71
Young Adults	28	55	83	1:1.96
All Subjects	86	96	182	1:1.12

3.1.3 Multiscale Representation of BOLD Signal

To capture neural correlates of reward processing at varying spatial and temporal scales, we can extract features from BOLD signals across multiple levels of resolution. This is shown in Figure 2. Anatomical regions are defined based on standard structural atlases, assuming consistent mapping between structure and function. Functionally defined ROIs are selected based on prior literature or task-related activations, under the assumption of functional homogeneity within each ROI. At the voxel and multi-voxel levels, we can use localized and pattern-based representations of the raw BOLD time series, which provide fine-grained spatial information but are more susceptible to noise. Temporally, we can analyze raw BOLD signals, trial-wise beta coefficients derived from event-related GLMs, and experiment-averaged beta values. These event-level and experiment-level beta estimates assume that the hemodynamic response function (HRF) Friston et al. (1998) is well-modeled and that neural responses are stationary across repeated trials. In this paper we have considered the beta values for each event and spatially aggregated them at the ROI level. This is identified by the dashed ellipse in Figure 2.

3.2 Model Methodology

The original Transformer architecture introduced a novel encoder-decoder framework composed entirely of attention mechanisms, achieving substantial improvements in machine translation tasks Vaswani et al. (2017). Since then, Transformer models have been widely adapted for classification and prediction tasks involving time-series data, particularly in the biomedical domain Katrompas et al. (2022); Nankani & Baruah (2022). These adaptations typically employ a Transformer encoder followed by a linear classification head to generate a probability distribution over output classes, following the approach introduced in the original BERT paper, which demonstrated the effectiveness of Transformers for classification tasks Devlin et al. (2018). In this study, we leverage the Transformer's ability to model sequential dependencies to predict high versus low impulsivity from trial-wise BOLD beta estimates. Specifically, we trained a Transformer model that incorporates both temporal and spatial embedding on trial-wise CUE and OUTCOME beta estimates to classify the impulsivity level of each subject.

3.2.1 Spatial and Temporal Embeddings

As Transformers process input sequences in parallel, they lack an inherent sense of order and therefore require positional embeddings to encode temporal context Vaswani et al. (2017). To align with the classification setup used in BERT Devlin et al. (2018), we prepend a special [CLS] token at the start of each subject's input sequence and insert [SEP] tokens to demarcate transitions between different experimental phases. Each input token is then represented as the sum of three types of embeddings: a float (value) embedding for the beta estimate, a regional embedding, which represent our eight ROIs, and a positional embedding, which captures trial-level sequence information by incrementing at each CUE and OUTCOME event[1]. Figure D-1 in the Appendix outlines the mapping of an example input trial sequence containing CUE and OUTCOME ROIs, first to its tokenized form and then to the subsequent positional and regional embeddings.

3.2.2 Model Architecture

We implemented a compact, single-layer Transformer architecture tailored for binary classification of impulsivity levels (high versus low) (Appendix E). The model architecture begins with a multi-head self-attention mechanism, which facilitates the extraction of temporal dependencies and contextual relationships within the input time series data. This is followed by a position-wise feed-forward network that applies non-linear transformations to each token independently, thereby enhancing the model's capacity to learn intricate feature representations. To ensure stable optimization and mitigate overfitting, each encoder sub-layer is followed by residual connections, layer normalization, and dropout regularization. The final encoder output is projected through a fully connected linear layer, with a softmax activation function applied to produce class probabilities corresponding to the binary impulsivity labels. This architectural design balances representational power with computational efficiency, making it well-suited for our neuroimaging dataset with a limited sample size. The model utilized eight attention heads and 128 hidden dimensions to effectively capture long-range dependencies inherent in the data, as each input sequence spanned the full duration of the experimental task.

3.2.3 Model Training

Model training was conducted using the Adam optimizer with a cross-entropy loss function. We divided the dataset into a 70/15/15 split for training, validation, and testing to assess model performance on unseen data. Hyperparameter optimization was performed within the training set to ensure generalizability across subjects. Specifically, we tuned the dropout rate and learning rate to mitigate overfitting, conducting separate hyperparameter searches for adolescents, young adults, and the combined sample to account for potential age-related differences in model performance. The training loop ran for up to 50 epochs, employing early stopping with a patience of 3 epochs based on validation loss to prevent overfitting. The model achieving the lowest validation loss was saved and restored for final evaluation. After training, the best model was evaluated on the held-out test set to generate final predictions and probabilities for downstream performance analyses. All computations were performed on a CPU.

Model performance was evaluated based on classification accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The ROC curve illustrates the trade-off between the true positive rate and the false positive rate across different classification thresholds Powers (2011). AUC-ROC was selected as a secondary evaluation metric due to its ability to capture the model's discriminative capacity in binary classification tasks, providing a more comprehensive assessment than accuracy alone Li (2024).

To improve model performance, we combined Gaussian noise augmentation with oversampling. Gaussian noise was added to the fMRI time series data Nguyen et al. (2020), where at each training step a random subset of sequences was perturbed by drawing values from a normal distribution and adding them element-wise to all non-special-token values. This simulated natural variability in the BOLD signal while preserving structural information, helping the model generalize and avoid overfitting. In addition, to address the class imbalance observed in our data, we applied the Synthetic Minority Oversampling Technique (SMOTE) Chawla et al. (2002), which has proven effective in

[1] We use the terms "spatial" and "regional" interchangeably, as well as the terms "temporal" and "positional."

small and imbalanced fMRI datasets Zhang et al. (2024). Both SMOTE and Gaussian noise augmentation were applied only to the training data and excluded from the validation and final test sets.

3.2.4 VISUALIZING THE ATTENTION MAPS

We evaluated our final three models, trained to classify high versus low impulsivity in adolescents, young adults, and all subjects, using attention maps. To interpret the learned representations, we extracted and visualized attention maps from each attention head, aggregated across all trials. Attention-based visualization has been widely used in neuroscience and biomedical domains to improve interpretability in deep learning models Vaswani et al. (2017); Abnar & Zuidema (2020); Nankani & Baruah (2022). The attention maps illustrate how the model distributes focus across different ROIs and task phases (tokens), offering insight into the spatio-temporal patterns relevant to impulsivity classification. This analysis aims to move beyond prediction accuracy toward interpretability, allowing us to probe the model's internal reasoning and potentially reveal neural mechanisms associated with impulsive behavior.

Other studies have explored pre-training on large-scale fMRI datasets, such as the Adolescent Brain Cognitive Development (ABCD) dataset, and then fine-tuning on task-specific data, to improve classification accuracy Kim et al. (2023); Kan et al. (2024). While these approaches have demonstrated promising performance gains, the primary focus of our work is on model interpretability and understanding the underlying neural mechanisms associated with reward processing and impulsivity. To maintain the validity of our results, we deliberately avoided pre-training on external datasets. Instead, we trained exclusively on our task-specific data to ensure that the learned representations remain directly tied to the specific neural processes of interest to our investigation.

4 RESULTS

In this section we provide a sample of the results. Additional results are provided in the Appendix submitted with supplementary material.

4.1 MODEL PERFORMANCE

Dropout rate and learning rate were selected separately for each subject group, with the results from hyperparameter tuning in Appendix F. We then trained models using 10 different random seeds for the train/validation/test split. The final model for attention map analysis was the one with the highest accuracy out of the 10 random seeds. Each model trained for up to 50 epochs, with early stopping applied if the validation loss failed to improve for three consecutive epochs. Across all models, early stopping occurred by epoch 11, indicating that the underlying representations were learned quickly.

Final model performance on the held-out test set is summarized in Table 2. The adolescent model achieved the highest test accuracy of 86.67%, indicating strong discriminative performance in distinguishing high versus low impulsivity within this group. The young adult model achieved a comparable best test accuracy of 84.62%, but had a much lower mean accuracy of 56.92%, compared with 66.59% across all adolescent models. The best model trained on all subjects achieved a test accuracy of 71.43%, suggesting that, despite the challenges of variability between age groups, the learned representations retain predictive ability across both adolescents and young adults.

Table 2: Model performance on the held-out test set (15% of the dataset), showing the distribution across 10 random seeds and the best model's result.

Subject Group	Mean Accuracy $\pm$ SD	Mean AUC $\pm$ SD	Best Accuracy	Best AUC
Adolescents	66.59 $\pm$ 10.89%	67.22 $\pm$ 10.97%	86.67%	90.74%
Young Adults	56.92 $\pm$ 16.57%	59.17 $\pm$ 17.22%	84.62%	80.56%
All Subjects	52.15 $\pm$ 10.86%	54.26 $\pm$ 10.13%	71.43%	70.77%

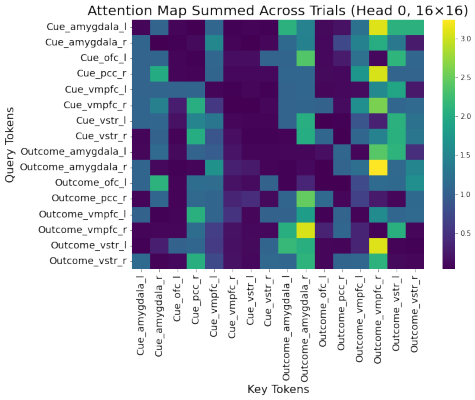

Figure 3: Attention map from Head 0 of the adolescent model, showing strong OUTCOME phase connectivity between the left ventral striatum and vmPFC, with additional cross phase links to vmPFC and amygdala activity, regions associated with emotional salience and decision making.

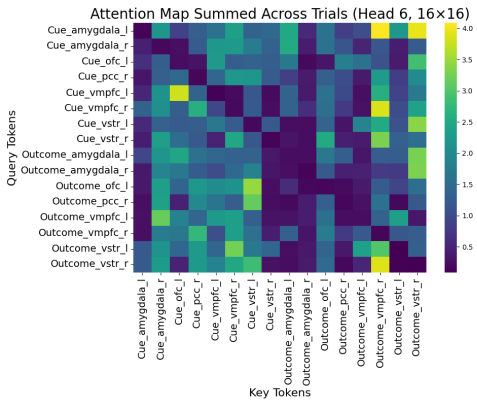

Figure 4: The attention map from Head 6 of the all subjects model highlights CUE→OUTCOME cross phase connectivity from the left amygdala to both the right vmPFC and the right ventral striatum, reflecting anticipatory signals that shape outcome evaluation.

4.2 Attention Maps

We generated attention maps for each of the final model's attention heads. Each attention head produces a matrix of size sequence length by sequence length, where each value is an attention score that represents how the token attends to every other token in the sequence. To focus on within-trial dynamics, we aggregated attention maps by isolating tokens belonging to the same trial, capturing how tokens within a trial attend to one another. This aggregation reflects the overall neural connectivity patterns represented in the data.

For visualization, we plotted attention values with key tokens on the X-axis and query tokens on the Y-axis. Key tokens represent the tokens being attended to (i.e., providing context), while query tokens represent the tokens attending to others (i.e., seeking context). Each model yields eight attention maps, corresponding to its eight attention heads. In total, 24 attention maps were generated across the three subject groups; here, we focus on two representative examples (Figures 3 for adolescents and 4 for all subjects), with the remaining maps provided in the Appendix G (see Figures G-1, G-2, and G-3). In the visualizations, higher attention weights are depicted in yellow, indicating stronger attention, while lower weights are shown in blue.

5 Discussion

Analyzing fMRI data to understand the links between reward processing and impulsivity in ADHD poses significant challenges due to a variety of factors. The data are inherently high-dimensional, with thousands of voxel-level time series per scan, creating a sparse subject-to-variable ratio and high variability across and within individuals. Clinical populations in general tend to produce more variable data, and this may be particularly the case during a dynamic developmental period such as adolescence and young adulthood when the brain regions studied here are undergoing significant change. Issues such as multicollinearity, lack of standardized acquisition methods, and limited interpretability of advanced models further complicate analysis Demirci et al. (2008); Specht (2020); Monti (2011); Grosenick et al. (2008). These challenges highlight the need for methods capable of modeling high-dimensional, variable data while remaining interpretable.

To address these challenges, we leveraged attention-based transformer models, which can capture complex spatio-temporal dependencies while providing interpretable insights. We analyzed task-based fMRI data collected using a Win/Loss paradigm, combined with impulsivity labels derived from standardized clinical measures combined with behavioral tasks associated with impulsivity and

reward sensitivity. Trial sequences from two conditions, CUE and OUTCOME, were encoded using positional and temporal embeddings to train a transformer encoder. The analysis was conducted separately for adolescents and young adults, and attention maps were extracted to identify the neural correlates of reward processing in individuals with low versus high impulsivity.

5.1 Observations on Model Performance

In addition to the class imbalance in impulsivity labels between young adults and adolescents mentioned in Section 3.1, differences in model performance between the adolescent and young adult groups in Table 2 may be related to developmental differences between the two stages of development. Sensitivity to reward and loss is hypothesized to be greater during the adolescent period Steinberg et al. (2018), and thus the imaging task we used may produce stronger results in an adolescent group. Furthermore, compensation in performance and neural processes may develop with maturity Fassbender & Schweitzer (2006) and compensatory processes may vary widely by the time young adults are maturing, producing more variability in the activity. Relatedly, adults may use a more cohesive network of brain regions to perform tasks with network connectivity stronger in the young adult period with less reliance on individual regions. Finally, differences in recruitment and referral to the study may have had an impact. The adolescents were primarily referred by their parents either for concerns about ADHD or as typically developing controls, whereas the young adults were largely self-referred. Self-referred young adults may represent a more heterogeneous population, with a broader range of symptom severity and self-perceived difficulties. This variability could attenuate group differences and reduce the model's ability to identify consistent predictive patterns.

Achieving over 70% best prediction accuracy across all subject groups is meaningful, as it suggests the model captures neural patterns associated with impulsivity that could inform understanding of clinically-elevated impulsivity versus more optimal self-control (i.e, low impulsivity) patterns in adolescents and young adults. Eventually, these findings could help illuminate the relation between neural data, behavior and clinical measures and inform future diagnostic or intervention strategies. Notably, the AUC-ROC values for all models are close to their corresponding accuracy values, indicating that the models maintain a consistent ability to differentiate between high and low impulsivity labels, even in the presence of class imbalance. This reflects the models' robustness in identifying relevant neural features associated with impulsivity-related behaviors.

5.2 Insights from Attention Maps

In the attention map for the adolescent model in Figure 3, the left ventral striatum (VStr) during the OUTCOME phase strongly attends to the left ventromedial prefrontal cortex (vmPFC) during the OUTCOME phase. Additional high-weight connections link the right Amygdala during the CUE and OUTCOME phases to the right vmPFC during the OUTCOME phase, and the left VStr during the CUE phase to the right Amygdala during the OUTCOME phase. These patterns suggest that classification of high versus low impulsivity was driven by coordinated engagement within the core reward and valuation network, with anticipatory valuation processes (VStr–vmPFC) playing a central role and cross-phase influences (amygdala→vmPFC, VStr→amygdala) integrating self-referential and emotional salience with reward receipt.

Importantly, we observe similar connectivity patterns—between the VStr, amygdala, and particularly the OUTCOME phase of the vmPFC—across the attention maps for the young adult and all subjects models, demonstrating that these findings are consistent across all subject groups and not restricted to adolescents (see Appendix Figures G-1, G-2, and G-3). During reward and loss tasks, the ventral striatum, vmPFC, and amygdala interact as core components of the frontostriatal-limbic circuitry that supports incentive-based learning and decision-making. In typical development, the VStr and vmPFC are strongly engaged during reward anticipation and receipt, supporting valuation and reinforcement learning, whereas the amygdala encodes the salience of both rewarding and aversive outcomes Oldham et al. (2018); Tang et al. (2024); Fareri et al. (2015). These roles align with our findings, which show that transformer models can reveal cross-temporal and cross-regional dependencies—such as interactions between anticipatory VStr, outcome-phase vmPFC, and amygdala signals—that are largely inaccessible to standard univariate fMRI analyses. This suggests that altered integration of anticipatory, valuation, and affective signals may contribute to heightened impulsivity in ADHD.

Our findings converge on a coherent neurobiological account of impulsivity that bridges adolescence and young adulthood. In adolescents, the transformer model highlighted ventral striatum–vmPFC coupling during reward anticipation, along with cross phase interactions involving the OFC and the amygdala, pointing to altered integration of anticipatory value signals, self referential processing, and affective salience in high impulsive individuals. Previous work has implicated heightened affective states, particularly irritability, in adolescents from this same study population Kahle et al. (2021); Komijani et al. (2025) with a latent class analysis suggesting a strong overlap between those with heightened irritability and impulsivity Elahi et al. (2024) and an association with atypical functional connectivity between the striatum and the amygdala Mukherjee et al. (2022). Crucially, these insights emerged from cross temporal and cross regional dependencies revealed by the transformer's attention mechanism, patterns that are largely inaccessible to conventional univariate fMRI analyses. By modeling how reward related signals evolve and interact over time, the approach captures subtle, clinically relevant neural features that could inform future diagnostic tools and personalized intervention strategies for impulsivity in ADHD.

6 Conclusion

In conclusion, we demonstrated that our transformer-based model effectively classifies subjects to high or low impulsivity from a task-based fMRI data. High accuracy was achieved when classifying adolescents, while accuracy for young adults was somewhat lower but still effective. Overall, these results suggest that the model is able to learn meaningful embeddings from the complex neural activations aggregated spatially at the ROI level and temporally for events of the reward processing task. We also visualized the attention maps that depict how different spatiotemporal activations (tokens) attend to each other. These attention maps broadly correlate to known brain activations during reward processing tasks. Overall, the study shows that event-based ROI data is a viable approach for capturing patterns relevant to impulsivity classification.

Future work will explore pre-training the model on larger datasets, including both internally collected data and publicly available resources such as the ABCD dataset Cohorts (2025). Additionally, incorporating auxiliary embeddings, such as trial-level win/loss probabilities or binary indicators reflecting whether a subject won or lost a specific trial, may further enhance the model's predictive ability. Finally, we intend to extend this framework to predict other clinically relevant symptoms such as irritability which is frequently observed in individuals with ADHD Karalunas et al. (2019); Leibenluft et al. (2024); Elahi et al. (2024).

7 Ethics Statement

The studies involving humans were approved by the institutional review board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Large Language Models (LLMs) assisted in editing grammar and clarity in certain sections of this paper.

8 Reproducibility Statement

Code and data will be made available upon acceptance of this paper to ICLR 2026. All randomness in training data augmentation and train/validation/test splits is controlled using fixed random seeds, which are specified in the code. Upon acceptance, we will provide an outline for reproducing our results using the provided code and data.

REFERENCES

Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. *CoRR*, abs/2005.00928, 2020. URL https://arxiv.org/abs/2005.00928.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, 5th ed., text rev. edition, 2022.

Nour-Mohammad Bakhshani. Impulsivity: a predisposition toward risky behaviors. *International Journal of High Risk Behaviors & Addiction*, 3(2):e20428, 2014. doi: 10.5812/ijhrba.20428.

Russell A. Barkley (ed.). *Attention-Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment*. The Guilford Press, New York, NY, 4th edition, 2015.

Russell A. Barkley. Neuropsychological testing is not useful in the diagnosis of adhd: Stop it (or prove it)! *The ADHD Report*, 27(2), 2019. URL https://guilfordjournals.com/doi/10.1521/adhd.2019.27.2.1.

Y. Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 5:157–66, 02 1994. doi: 10.1109/72.279181.

Danilo Bzdok. Classical statistics and statistical learning in imaging neuroscience. *Frontiers in Neuroscience*, 11:543, 2017. doi: 10.3389/fnins.2017.00543.

Henry W. Chase, Jay C. Fournier, Michael A. Bertocci, Tamar Greenberg, Hina Aslam, Rachel Stiffler, John Lockovich, Silvia Graur, Genna Bebko, Erika E. Forbes, and Mary L. Phillips. A pathway linking reward circuitry, impulsive sensation-seeking and risky decision-making in young adults: identifying neural markers for new interventions. *Translational Psychiatry*, 7(4): e1096, April 2017. doi: 10.1038/tp.2017.60.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

NIH Brain Development Cohorts. Adolescent brain cognitive development (abcd) study, 2025. URL https://www.nbdc-datahub.org/abcd-study.

Shan Cong, Hang Wang, Yang Zhou, Zheng Wang, Xiaohui Yao, and Chunsheng Yang. Comprehensive review of transformer-based models in neuroscience, neurology, and psychiatry. *Brain-X*, 2:e57, 2024. doi: 10.1002/brx2.57.

Robert W. Cox. Afni: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162–173, 1996. doi: 10.1006/cbmr.1996.0014.

R. Cameron Craddock, G. Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012. doi: 10.1002/hbm.21333.

Peishan Dai, Ying Zhou, Yun Shi, Da Lu, Zailiang Chen, Beiji Zou, Kun Liu, Shenghui Liao, and The REST meta MDD Consortium. Classification of mdd using a transformer classifier with large-scale multisite resting-state fmri data. *Human Brain Mapping*, 45(1):e26542, 2024. doi: 10.1002/hbm.26542.

Kimberlee D'Ardenne, Samuel M. McClure, Leigh E. Nystrom, and Jonathan D. Cohen. Bold responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, 319 (5867):1264–1267, 2008. doi: 10.1126/science.1150605.

Oguz Demirci, Vincent P. Clark, Vincent A. Magnotta, Nancy C. Andreasen, John Lauriello, Kent A. Kiehl, Godfrey D. Pearlson, and Vince D. Calhoun. A review of challenges in the use of fmri for disease classification/characterization and a projection pursuit application from a multi-site fmri schizophrenia study. *Brain Imaging and Behavior*, 2:207–226, 2008.

Xin Deng, Jiahao Zhang, Rui Liu, and Ke Liu. Classifying asd based on time-series fmri using spatial–temporal transformer. *Computers in Biology and Medicine*, 151, 2022. doi: 10.1016/j.compbiomed.2022.106320.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Heather Elahi, Ana-Maria Iosif, Prerona Mukherjee, Stephen P. Hinshaw, and Julie B. Schweitzer. Using hot and cool measures to phenotype and predict functional outcomes across dimensions of adhd and typical development in adolescents. *Research on Child and Adolescent Psychopathology*, 52(4):579–593, 2024. doi: 10.1007/s10802-023-01023-7.

Blake L. Elliott, Kimberlee D'Ardenne, Prerona Mukherjee, Julie B. Schweitzer, and Samuel M. McClure. Limbic and executive meso- and nigrostriatal tracts predict impulsivity differences in attention-deficit/hyperactivity disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 7(4):415–423, 2022. doi: 10.1016/j.bpsc.2021.10.002.

Stephen V. Faraone, Tobias Banaschewski, David Coghill, Yi Zheng, Joseph Biederman, Mark A. Bellgrove, Jeffrey H. Newcorn, Monique Gignac, Nawaf M. Al Saud, Iris Manor, Luis A. Rohde, Lin Yang, Samuele Cortese, Dvora Almagor, Mark A. Stein, Thamer H. Albatti, Huda F. Aljoudi, Mohammed M. J. Alqahtani, Philip Asherson, Lukoye Atwoli, Sven Bölte, Jan K. Buitelaar, Chris L. Crunelle, David Daley, Sune Dalsgaard, Manfred Döpfner, Sandra Espinet, Michelle Fitzgerald, Barbara Franke, Matthias Gerlach, Janette Haavik, Catharina A. Hartman, Christina M. Hartung, Stephen P. Hinshaw, Pieter J. Hoekstra, Chris Hollis, Scott H. Kollins, J. J. Sandra Kooij, Jonna Kuntsi, Henrik Larsson, Tian Li, Jianghong Liu, Eli Merzon, Grace Mattingly, Paula Mattos, Scott McCarthy, Amori Y. Mikami, Brooke S. G. Molina, Joel T. Nigg, Diane Purper-Ouakil, Olayinka O. Omigbodun, Guilherme V. Polanczyk, Yair Pollak, Alison S. Poulton, Ravi P. Rajkumar, Andrew Reding, Andreas Reif, Katya Rubia, Julia Rucklidge, Marcel Romanos, Jose A. Ramos-Quiroga, Anouk Schellekens, Anouk Scheres, Riaan Schoeman, Julie B. Schweitzer, Humera Shah, Mary V. Solanto, Edmund Sonuga-Barke, Carlos Soutullo, Hans-Christoph Steinhausen, James M. Swanson, Anita Thapar, Gail Tripp, Gerdien van de Glind, Wim van den Brink, Saskia Van der Oord, Andries Venter, Benedetto Vitiello, Susanne Walitza, and Yi Wang. The world federation of adhd international consensus statement: 208 evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews*, 128:789–818, September 2021. doi: 10.1016/j.neubiorev.2021.01.022. Epub 2021 Feb 4.

Dominic S. Fareri, Laurel Gabard-Durnam, Bonnie Goff, Jessica Flannery, Dylan G. Gee, Daniel S. Lumian, Christina Caldera, and Nim Tottenham. Normative development of ventral striatal resting state connectivity in humans. *NeuroImage*, 118:422–437, September 2015. doi: 10.1016/j.neuroimage.2015.06.022. Epub 2015 Jun 16.

Catherine Fassbender and Julie B. Schweitzer. Is there evidence for neural compensation in attention deficit hyperactivity disorder? a review of the functional neuroimaging literature. *Clinical Psychology Review*, 26(4):445–465, August 2006. doi: 10.1016/j.cpr.2006.01.003. Epub 2006 Feb 24.

Karl J. Friston, P. Fletcher, Oliver Josephs, Andrew Holmes, M. D. Rugg, and Robert Turner. Event-related fmri: characterizing differential responses. *Neuroimage*, 7(1):30–40, 1998.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey. *arXiv preprint arXiv:2101.00734*, 2021.

Logan Grosenick, Stephanie Greer, and Brian Knutson. Interpretable classifiers for fmri improve prediction of purchases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(6):539–548, 2008.

Stephen P. Hinshaw, Elizabeth B. Owens, Catherine Zalecki, Stephanie P. Huggins, Ana J. Montenegro-Nevado, Elizabeth Schrodek, and Elizabeth N. Swanson. Prospective follow-up of girls with attention-deficit/hyperactivity disorder into early adulthood: continuing impairment includes elevated risk for suicide attempts and self-injury. *Journal of Consulting and Clinical Psychology*, 80(6):1041–1051, December 2012. doi: 10.1037/a0029451. Epub 2012 Aug 13.

Xiaojie Huang, Jun Xiao, and Chao Wu. Design of deep learning model for task-evoked fmri data classification. *Computational Intelligence and Neuroscience*, 2021:6660866, 2021. doi: 10.1155/2021/6660866.

Sarah Kahle, Prerona Mukherjee, J. Faye Dixon, Ellen Leibenluft, Stephen P. Hinshaw, and Julie B. Schweitzer. Irritability predicts hyperactive/impulsive symptoms across adolescence for females. *Research on Child and Adolescent Psychopathology*, 49(2):185–196, 2021. doi: 10.1007/s10802-020-00721-3.

Xuan Kan, Hejie Cui, Keqi Han, Ying Guo, and Carl Yang. Multi-task learning for brain network analysis in the abcd study. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–8, 2024. doi: 10.1109/BHI62660.2024.10913627.

Sarah L. Karalunas, Hanna C. Gustafsson, Damien Fair, Erica D. Musser, and Joel T. Nigg. Do we need an irritable subtype of ADHD? replication and extension of a promising temperament profile approach to ADHD subtyping. *Psychological Assessment*, 31(2):236–247, February 2019. doi: 10.1037/pas0000664. Epub 2018 Oct 25.

Alexander Katrompas, Theodoros Ntakouris, and Vangelis Metsis. Recurrence and self-attention vs the transformer for time-series classification: A comparative study. In Martin Michalowski, Syed Sibte Raza Abidi, and Samina Abidi (eds.), *Artificial Intelligence in Medicine*, pp. 99–109, Cham, 2022. Springer International Publishing. ISBN 978-3-031-09342-5.

Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer, 2023. URL `https://arxiv.org/abs/2307.05916`.

Kenneth N. Kirby, Nancy M. Petry, and Warren K. Bickel. Heroin addicts have higher discount rates for delayed rewards than non–drug-using controls. *Journal of Experimental Psychology: General*, 128(1):78–87, 1999. doi: 10.1037/0096-3445.128.1.78.

Saeedeh Komijani, Dipak Ghosal, Manpreet K. Singh, Julie B. Schweitzer, and Prerona Mukherjee. A novel framework to predict adhd symptoms using irritability in adolescents and young adults with and without adhd. *Frontiers in Psychiatry*, 15:1467486, 2025. doi: 10.3389/fpsyt.2024.1467486.

Ellen Leibenluft, Laura E. Allen, Robert R. Althoff, Melissa A. Brotman, Jeffrey D. Burke, Gabrielle A. Carlson, Daniel P. Dickstein, Lea R. Dougherty, Spencer C. Evans, Kathryn Kircanski, Daniel N. Klein, Edward P. Malone, Carla A. Mazefsky, Joel Nigg, Susan B. Perlman, Daniel S. Pine, Amy Krain Roy, Giovanni A. Salum, Aaron Shakeshaft, Jessica Silver, Joel Stoddard, Anita Thapar, Wan-Ling Tseng, Pablo Vidal-Ribas, Lauren S. Wakschlag, and Argyris Stringaris. Irritability in youths: A critical integrative review. *American Journal of Psychiatry*, 181(4):275–290, April 2024. doi: 10.1176/appi.ajp.20230256. Epub 2024 Feb 29.

Carl W. Lejuez, Jennifer P. Read, Christopher W. Kahler, Jessica B. Richards, Suzanne E. Ramsey, Gregory L. Stuart, David R. Strong, and Richard A. Brown. Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, 8(2):75–84, 2002. doi: 10.1037/1076-898X.8.2.75.

Jung Li. Area under the roc curve has the most consistent evaluation for binary classification. *PLoS ONE*, 19(12), 2024.

Matthew D. Lieberman, Elliot T. Berkman, and Tor D. Wager. Correlations in social neuroscience aren't voodoo. *Perspectives on Psychological Science*, 4(3):299–302, 2009. doi: 10.1111/j.1745-6924.2009.01128.x.

Martin M. Monti. Statistical analysis of fmri time-series: a critical review of the glm approach. *Frontiers in Human Neuroscience*, 5:28, 2011. doi: 10.3389/fnhum.2011.00028.

Prerona Mukherjee, Tadeus Hartanto, Ana-Maria Iosif, J. Faye Dixon, Stephen P. Hinshaw, Murat Pakyurek, Wouter van den Bos, et al. Neural basis of working memory in adhd: Load versus complexity. *NeuroImage: Clinical*, 30:102662, 2021. doi: 10.1016/j.nicl.2021.102662.

Prerona Mukherjee, Veronika Vilgis, Shawn Rhoads, Rajpreet Chahal, Catherine Fassbender, Ellen Leibenluft, J. Faye Dixon, et al. Associations of irritability with functional connectivity of amygdala and nucleus accumbens in adolescents and young adults with adhd. *Journal of Attention Disorders*, 26(7):1040–1050, 2022. doi: 10.1177/10870547211057074.

Lindsay Munroe, Mariana da Silva, Faezeh Heidari, Irina Grigorescu, Simon Dahan, Emma C. Robinson, Maria Deprez, and Po-Wah So. Applications of interpretable deep learning in neuroimaging: A comprehensive review. *Imaging Neuroscience*, 2, 2024. doi: 10.1162/imag_a_00214.

Deepankar Nankani and Rashmi Dutta Baruah. Atrial fibrillation classification and prediction explanation using transformer neural network. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 01–08, 2022. doi: 10.1109/IJCNN55064.2022.9892286.

Kevin P. Nguyen, Cherise Chin Fatt, Alex Treacher, Cooper Mellema, Madhukar H. Trivedi, and Albert Montillo. Anatomically-informed data augmentation for functional mri with applications to deep learning. In *Proceedings of the SPIE International Society for Optics and Photonics*, volume 11313, 2020. doi: 10.1117/12.2548630. Available in PMC 2021 March 24.

Sara M. O'Grady and Stephen P. Hinshaw. Long-term outcomes of females with attention-deficit hyperactivity disorder: increased risk for self-harm. *British Journal of Psychiatry*, 218(1):4–6, January 2021. doi: 10.1192/bjp.2020.153.

Stuart Oldham, Carsten Murawski, Alex Fornito, George Youssef, Murat Yücel, and Valentina Lorenzetti. The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Human Brain Mapping*, 39(8):3398–3418, 2018. doi: 10.1002/hbm.24184.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/pascanu13.html.

Jim H. Patton, Matthew S. Stanford, and Ernest S. Barratt. Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, 51(6):768–774, 1995. doi: 10.1002/1097-4679(199511)51:6⟨768::AID-JCLP2270510607⟩3.0.CO;2-1.

Joseph M. Pierre. *Overdiagnosis, underdiagnosis, synthesis: A dialectic for psychiatry and the DSM*, chapter 8. Springer New York, 2013.

Russell A. Poldrack. Region of interest analysis for fmri. *Social Cognitive and Affective Neuroscience*, 2(1):67–70, 2007. doi: 10.1093/scan/nsm006.

David M. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Romy Sauvayre. "voodoo" science in neuroimaging: How a controversy transformed into a crisis. *Social Sciences*, 12(1):15, 2023. doi: 10.3390/socsci12010015.

Karsten Specht. Current challenges in translational and clinical fmri and future directions. *Frontiers in Psychiatry*, 10:924, 2020.

Laurence Steinberg, Grace Icenogle, Elizabeth P. Shulman, Kaitlyn Breiner, Jason Chein, Dario Bacchini, Lei Chang, Nandita Chaudhary, Laura Di Giunta, Kenneth A. Dodge, Kostas A. Fanti, Jennifer E. Lansford, Patrick S. Malone, Paul Oburu, Concetta Pastorelli, Ann T. Skinner, Emma Sorbring, Sombat Tapanya, Liliana M. U. Tirado, Liane P. Alampay, Suha M. Al-Hassan, and Hoda M. S. Takash. Around the world, adolescence is a time of heightened sensation seeking and immature self-regulation. *Developmental Science*, 21(2):e12532, March 2018. doi: 10.1111/desc.12532. Epub 2017 Feb 1.

Hua Tang, Romane Bartolo, and Bruno B. Averbeck. Ventral frontostriatal circuitry mediates the computation of reinforcement from symbolic gains and losses. *Neuron*, 112(22):3782–3795.e5, November 2024. doi: 10.1016/j.neuron.2024.08.018. Epub 2024 Sep 24.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3):274–290, 2009. doi: 10.1111/j.1745-6924.2009.01125.x.

Wei Zhang, Wei Zeng, Hui Chen, Jun Liu, Hongbo Yan, Kai Zhang, Rui Tao, Wai Ting Siok, and Ning Wang. Stanet: A novel spatio-temporal aggregation network for depression classification with small and unbalanced fmri data. *Tomography*, 10(12):1895–1914, 2024. doi: 10.3390/tomography10120138. Published 2024-11-28.

Hongtu Zhu, Tengfei Li, and Bingxin Zhao. Statistical learning methods for neuroimaging data analysis with applications. *Annual Review of Biomedical Data Science*, 6:73–104, 2023. doi: 10.1146/annurev-biodatasci-020722-100353.

Philip G. Zimbardo and John N. Boyd. Putting time in perspective: A valid, reliable individual-differences metric. *Journal of Personality and Social Psychology*, 77(6):1271–1288, 1999. doi: 10.1037/0022-3514.77.6.1271.

APPENDIX

A DATA ACQUISITION AND PREPROCESSING

The imaging data were acquired using a Siemens 3T TIM Trio MRI scanner equipped with a 32-channel head coil. Functional T2* images were collected with a voxel size of 3.4 mm × 3.4 mm × 3.4 mm, slice thickness of 3.4 mm (isotropic), across 36 interleaved slices. The repetition time (TR) was 2.0 s, the echo time (TE) was 25 ms, the flip angle was 90°, and the matrix size was 64 × 64 with a field of view (FOV) of 220 mm. The task-based fMRI included four runs, each consisting of 182 volumes. Additionally, high-resolution anatomical images were acquired using an MPRAGE sequence (TR = 1.9 s, TE = 3.06 ms, FOV = 256 mm, matrix = 256 × 256, flip angle = 7°, slice thickness = 1 mm, 208 slices). The experimental stimuli were presented using E-Prime 2.0 software.

The fMRI data were analyzed using Analysis of Functional NeuroImages(AFNI) tool Cox (1996). The initial two volumes from each scan were discarded to allow for signal stabilization. Preprocessing involved removing non-brain tissue, followed by aligning each run to the participant's T1-weighted structural MRI and transforming the data into Montreal Neurological Institute (MNI) space. Registration was performed using FMRIB's Linear Image Registration Tool . Smoothing was applied with a 4 mm full-width at half-maximum (FWHM) Gaussian filter, and normalization was conducted in accordance with our previous studies Mukherjee et al. (2021). The voxel size was resampled to 2 mm^3. Volumes exhibiting motion exceeding 1 mm between successive scans were excluded from further analysis. Participants with more than %25 of their volumes omitted due to motion were excluded from the study.

B BRAIN REGIONS OF INTEREST (ROI) MAPPING

Table B1 presents the full names of the brain regions analyzed in this study—selected based on their role in reward processing—together with the shorthand codes used in the dataset and their corresponding numeric identifiers.

Table B1: Mapping of brain regions to shorthand and numeric IDs

ROI Full Name	**Shorthand**	**ID**
Left Amygdala (AMYG)	amygdala_l	1
Right Amygdala (AMYG)	amygdala_r	2
Left Orbitofrontal Cortex (OFC)	ofc_l	3
Right Posterior Cingulate Cortex (PCC)	pcc_r	4
Left Ventromedial Prefrontal Cortex (vmPFC)	vmpfc_l	5
Right Ventromedial Prefrontal Cortex (vmPFC)	vmpfc_r	6
Left Ventral Sriatum (VStr)	vstr_l	7
Right Ventral Striatum (VStr)	vstr_r	8

C FEATURE SPACE DIMENSIONALITY

This section provides an overview of the dimensions of the resulting input sequences for the transformer model. The input feature space used in this study is summarized in Table C1. Features were constructed by extracting scalar brain activation values from each region of interest (ROI) across task-related timepoints.

D SPATIAL AND TEMPORAL EMBEDDING EXAMPLE

Figure D-1 denotes the spatial and temporal embedding for the first trial. The input sequence is embedded by summing three components: tokenized input embeddings, a regional embedding encoding spatial brain ROI information, and a positional embedding that encodes CUE/OUTCOME event phases.

Table C1: Input feature space dimensionality per subject

Feature Type	**Dimension / Size**
Number of ROIs	8
Number of phases per trial	2 (CUE and OUTCOME)
Number of trials per subject	36 (min) to 60 (max)
Total input features per subject	576 (min) to 960 (max)

Input Embedding	'CLS'	CUE (ROI1)	...	CUE (ROI8)	'SEP'	OUTCOME (ROI1)	...	OUTCOME (ROI8)	'SEP'
Regional Embedding	0	1	...	8	0	1	...	8	0
Positional Embedding	0	1	...	1	0	2	...	2	0

Figure D-1: Example embedding for the first trial. The positional embedding increments with each subsequent CUE/OUTCOME event, reflecting the temporal nature of task phases within a trial; each trial consists of a CUE event followed by an OUTCOME event at the next time point.

E Overview of the Transformer Model

Figure E-1 provides an overview of the transformer-based architecture used in our study. Each input sequence consists of scalar brain activation values from the CUE and OUTCOME phases for each subject. These values are embedded using three summed embeddings (input, regional, and positional; see Figure D-1). The resulting sequence is passed through a Transformer encoder, which applies multi-head self-attention and a feed-forward network. The encoded representation is then processed through a linear layer followed by a softmax function to generate a binary output classifying individuals as low versus high in impulsivity.

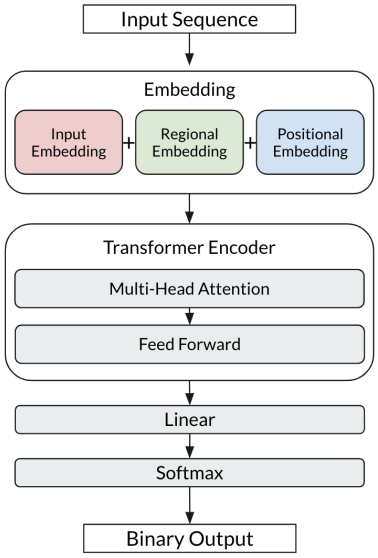

Figure E-1: Transformer-based model architecture.

F Hyperparameter Tuning Performance Results

Table F1 presents a comparison of model performance across different dropout rates and learning rates for adolescents, young adults, and the combined sample. Both accuracy and AUC-ROC metrics

are reported for each group. For adolescents, the best performance was achieved with a dropout rate of 0.2 and a learning rate of 10^{-4}, yielding an accuracy of 0.8000 and an AUC-ROC of 0.7778. In young adults, performance was more modest, with the highest accuracy (0.6154) and AUC-ROC (0.5556) obtained at a dropout rate of 0.3 and a learning rate of 10^{-3}. For the combined sample, the optimal setting was a dropout rate of 0.3 and a learning rate of 10^{-4}, resulting in an accuracy of 0.6786 and an AUC-ROC of 0.7128. Bolded values in the table highlight these best-performing hyperparameter configurations, which were used to guide to train the final models for each group.

Table F1: Comparison of test accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC, or simply AUC) for adolescents, young adults, and combined subjects across dropout rates and learning rates. Bolded values indicate the best-performing settings chosen for each group and metric.

Dropout Rate	Learning Rate	Adolescents		Young Adults		Combined Subjects	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
0.2	10^{-3}	0.6000	0.7593	0.6154	0.5556	0.5357	0.7282
0.2	10^{-4}	**0.8000**	**0.7778**	0.4615	0.5556	0.5000	0.6718
0.3	10^{-3}	0.6000	0.8333	**0.6154**	**0.5556**	0.4643	0.7077
0.3	10^{-4}	0.6667	0.7778	0.5385	0.5278	**0.6786**	**0.7128**
0.4	10^{-3}	0.6000	0.8333	0.3077	0.6389	0.5357	0.7282
0.4	10^{-4}	0.7333	0.8148	0.4615	0.5556	0.6071	0.6103

G Attention Maps

We visualize the attention maps of our 8 attention heads for each model including adolescents, young adults, and all subjects. Attention scores are aggregated across trials, showing the interaction patterns between spatiotemporal tokens within individual trials.

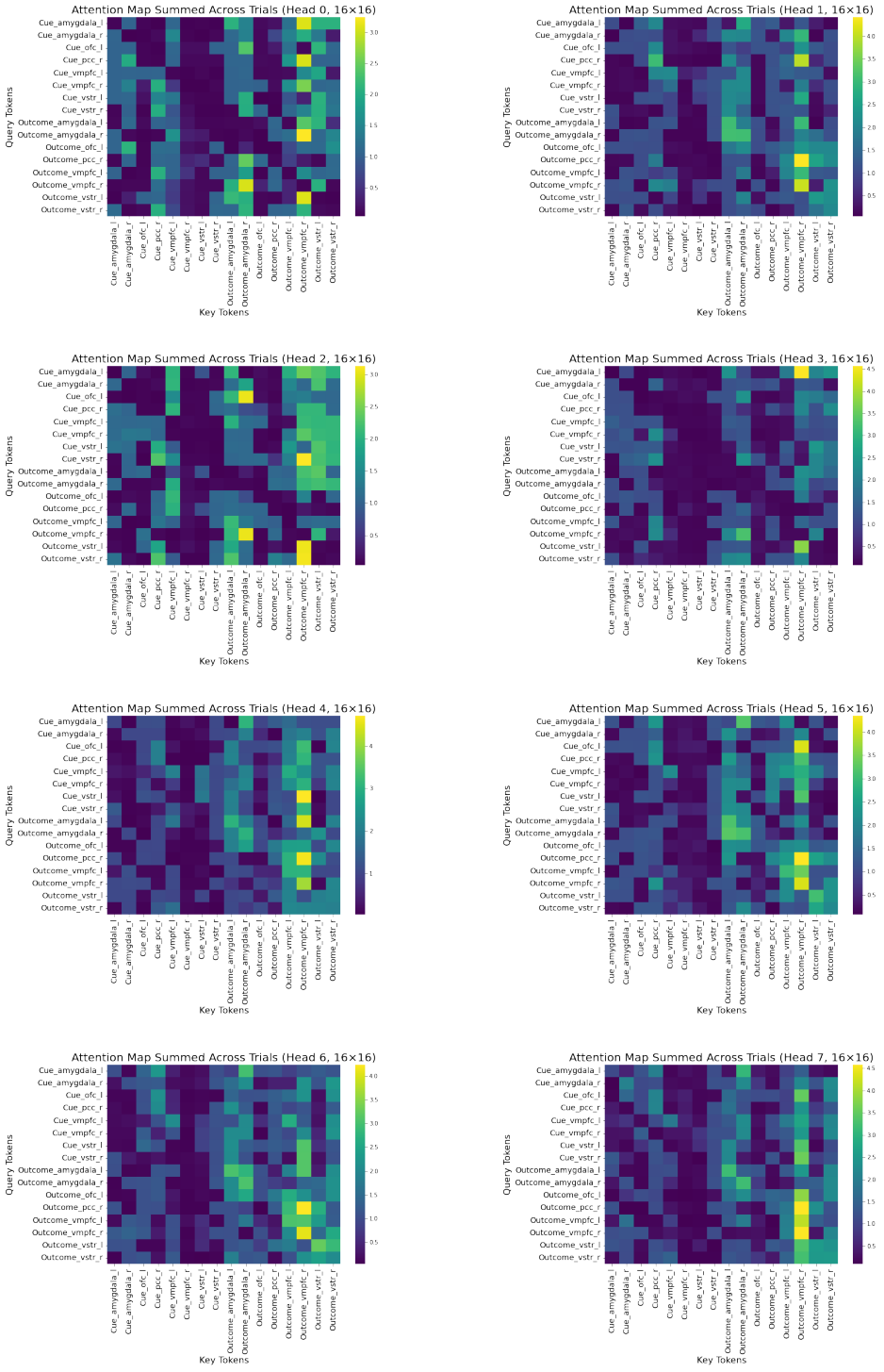

Figure G-1: Attention maps for the eight attention heads (numbered 0–7, left to right and top to bottom) show cross-phase and inter-phase OUTCOME interactions between the amygdala, ventral striatum (VStr), PCC, and vmPFC in the adolescent model.

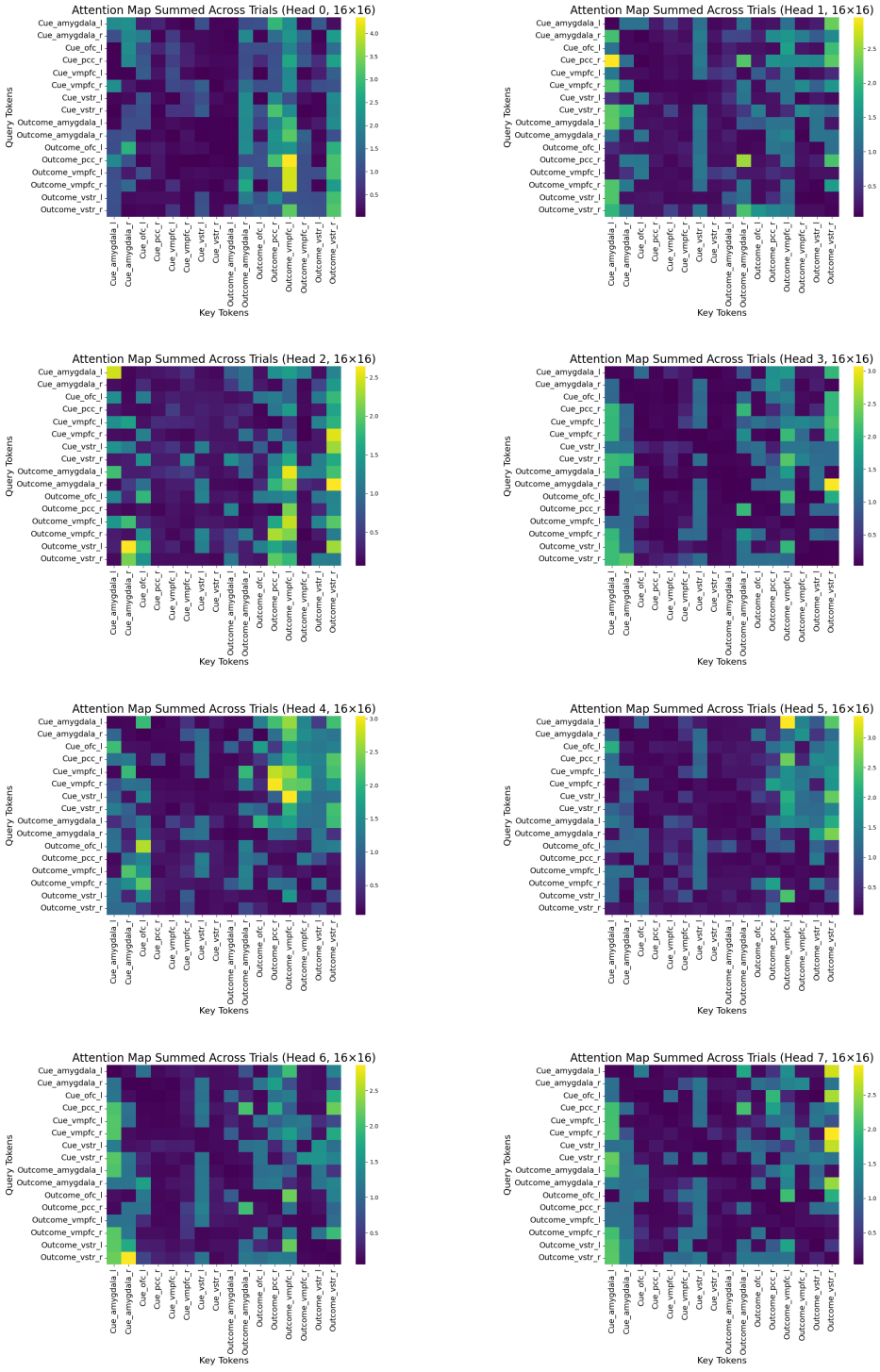

Figure G-2: Attention maps for the eight attention heads (numbered 0–7, left to right and top to bottom) in the young adult model highlight show cross-phase interactions between the amygdala, ventral striatum (VStr), PCC, and vmPFC.

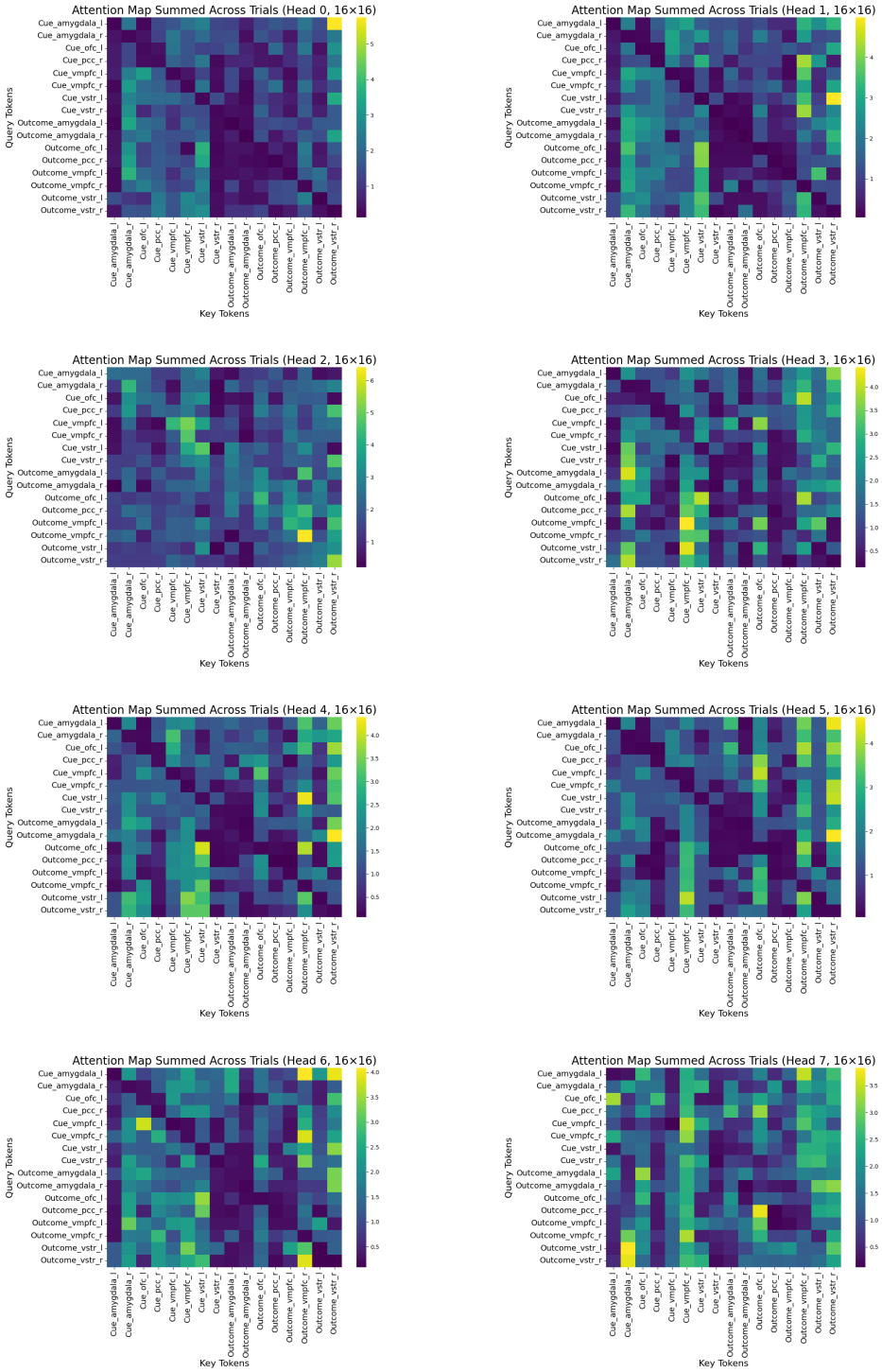

Figure G-3: Attention maps for the eight attention heads (numbered 0–7, left to right and top to bottom) in the all subjects model show, in particular, that heads 0, 5, and 6 reveal strong connectivity between the left amygdala during the CUE phase and the right ventral striatum (VStr) during the OUTCOME phase, reflecting the link between emotional salience and reward processing.