Appendix

Organization of the Appendix

In Appendix A, we provide a detailed description on how the eye-hand behavioral sequences are collected, including the VR setup and how fixations are extracted. We also provide precise coordinates for the boundaries of different regions on the screen. We demonstrate the irregularities and challenges in the data in Appendix B, serving as the motivation of our model design. In Appendix C, we provide additional information about the internal structure of the eye-to-hand translation model, along with the parameters used across the experiments. In Appendix D, we provide several extensions of our results, along with visualizations. We also compare the effectiveness of our method with an alignment method in Appendix D, proving that aligning behavioral sequences from different modalities is less effectively, which may actually lead to worse results.

A Detailed Description of the Dataset

The data was collected through a Multimodal Virtual Classroom Interface (MVCI) system (Zhiwei Yu 2023), with layout visualized in Figure 1. IRB approval has been received before the data collection experiments were conducted. The system simulates an interactive classroom environment, with an interactive screen that includes a simple drawing task for the individual. The classroom incorporated auditory, visual, and tactile stimuli, such as being able to hear the avatar moving around or a bell ringing. The participant was given a stylus and could receive touch feedback from it, such as friction, weight, drag, and resistance, as if they were holding an actual pen. Additionally, the tasks require important skills in a classroom that need continuous coordination between hand and eye movements. The system applied the dispersion threshold algorithm for fixation identification (I-DT). For eye gaze, the angular threshold was $0.7°$, and the minimum duration threshold was 275ms. For hand motion, the angular threshold was $50°$, and the minimum duration threshold was 1000ms.

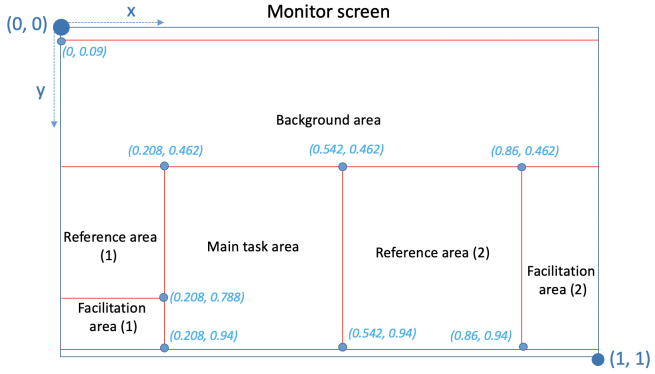

Figure 2: Coordinate boundaries of different regions on the screen which correspond to different tasks or images on the screen.

All participants were between the ages of 11 to 17, and we collected data from 9 autistic adolescents and 17 TD adolescents. The regions boundaries correspond to the activities on the screen, and their coordinates are shown in Figure 2.

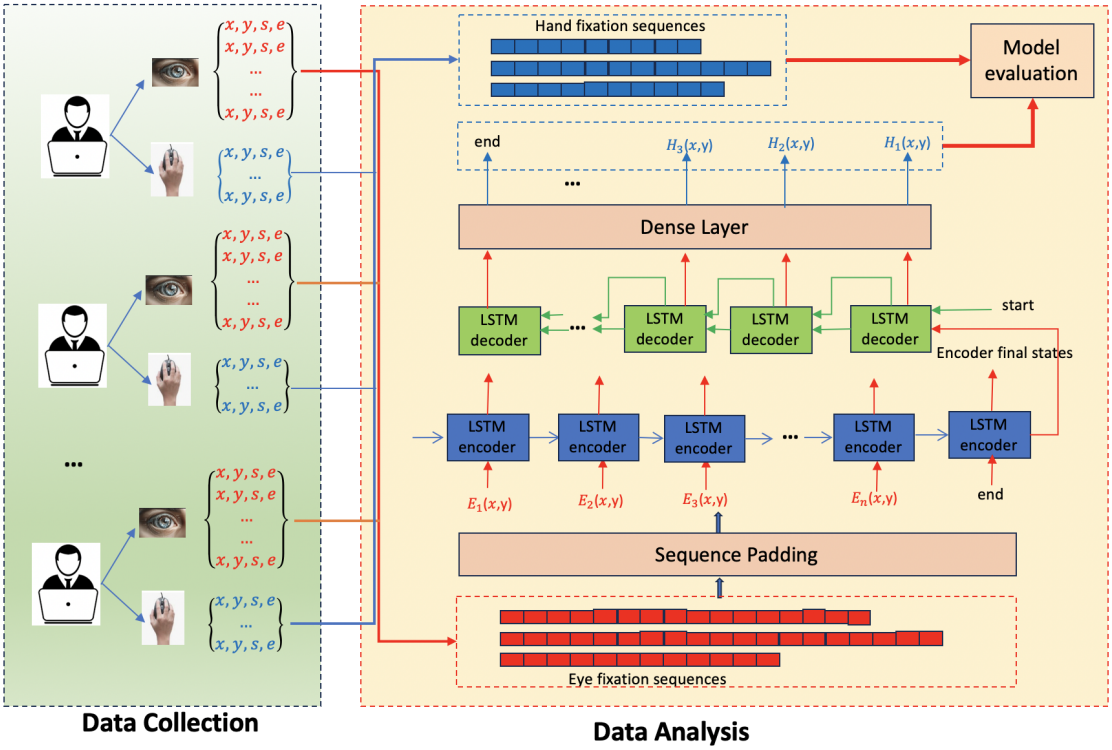

Figure 4: A Multi-modal translation framework for eye-hand coordination Analysis

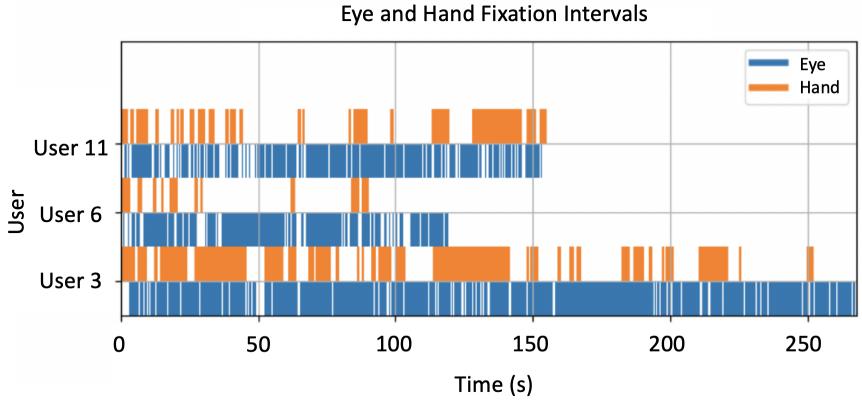

Figure 3: Visualization of fixation sequences, which are highly irregular and noisy, being of different lengths and are not aligned. The data varies largely from user to user, with both very concentrated data and very sparse data in the same sequence.

B Additional Details of the Challenges of the Problem

To further demonstrate the challenges in analyzing the multimodal behavioral data, Figure 3 uses a bar to represent each time interval, where there is a fixation. It shows how the hand and eye fixations are highly noisy and unaligned, and the sequence lengths are largely varied, where users such as User 3 have a very long fixation sequence, while User 6 has a relatively short fixation sequence. Fixation lengths are also highly varied, with some being almost 50 seconds, as seen in User 3, and others being very short and concentrated as in the start of User 11's sequence. Additionally, User 6's sequence starts out very concentrated, with multiple eye fixations in the first 30 seconds, but later in the sequence, these fixations become sparse. We

see how the eye and hand fixations happen at largely different times with varying irregularity, making it difficult to pinpoint exactly which eye fixation corresponds to each hand fixation.

C Additional Details of the Methodology

Using LSTM as an example, our model adopts an encoder-decoder structure to perform the eye-to-hand translation. The eye fixation sequences are extracted from the collected data and used as input for the LSTM encoder. An LSTM is typically used in an encoder-decoder structure to capture long-term contexts within an input sequence. Due to the variable lengths of sequences, we first applied the sequence padding to ensure the sequences can be handled properly and efficiently by the model. As shown in Figure 4, we have input and output sequences of eye and hand fixations, respectively, and our model tries to predict the output sequence based on the input sequence. The decoder cell takes the final hidden cell states from the encoder as its initial states. Using these values, we generate the output sequences by taking in y_{t-1} along with h_{t-1} and c_{t-1} and continuously update h and c. It then uses the current hidden states to predict y_t (which is produced by a dense layer). All experiments were run with latent dims set to 64 and for $1,000$ epochs. Figure 5 demonstrates both autistic and TD groups converge by $1,000$ epochs.

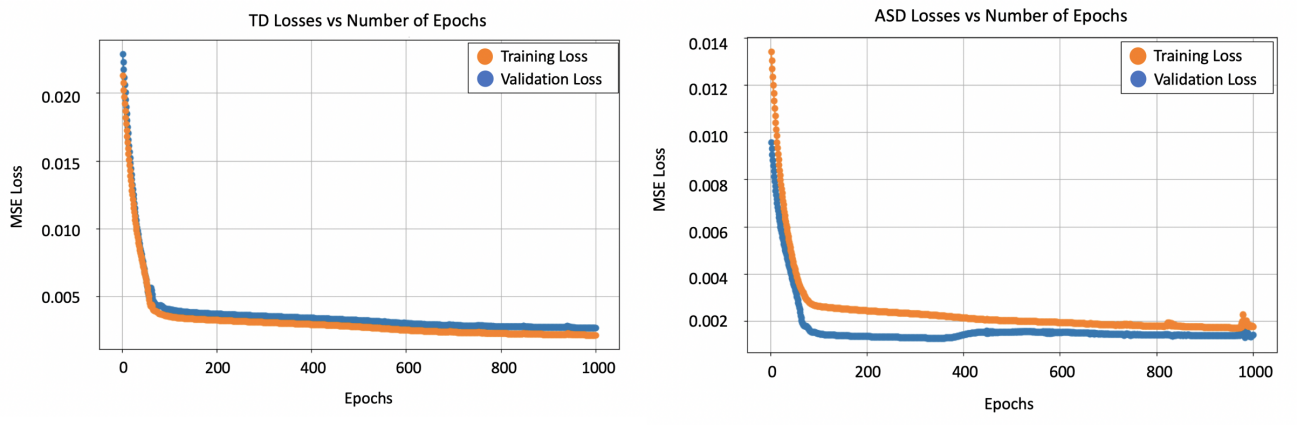

Figure 5: Training and validation losses over the training epochs, which show convergence for both the autistic and TD curves.

D Additional Experimental Results

D.1 Additional Results from the Eye-To-Hand Translation Models

In Table 3, we include several more comparisons between the models used. The numbers in the intersection of a row and a column contain both the T-statistic and P-value of the two sets of validation errors corresponding to the two methods. Bold entries have P-value < 0.05, indicating a statistically significant difference. We see that a difference in the number of features also has an effect on the prediction error. In particular, it seems that including the time of fixations doesn't seem to be a big factor in prediction, giving significantly worse predictions in most cases as seen in Table 1. For example, the LSTM on the TD group, has a much larger mean error with time as a feature and a P-value of 0.0004585 compared to using position as a feature.

Table 3: Statistical significance of the comparison results between models on both autistic and TD behavioral data

	Aut (T-Statistic/P-value)		**TD** (T-Statistic/P-value)	
Model	GRU-P	LSTM-PT	GRU-P	LSTM-PT
GRU-PT	**-2.1645/0.009640**	-0.06111/0.9542	-1.745/0.1559	**12.073/0.0002700**
LSTM-P	0.8736/0.4316	-1.352/0.2477	**5.794/0.004410**	**-10.5606/0.0004585**

We also provide a visualization in Figure 6 of the results of the region-based method. We see that the TD errors are always higher than the autistic errors.

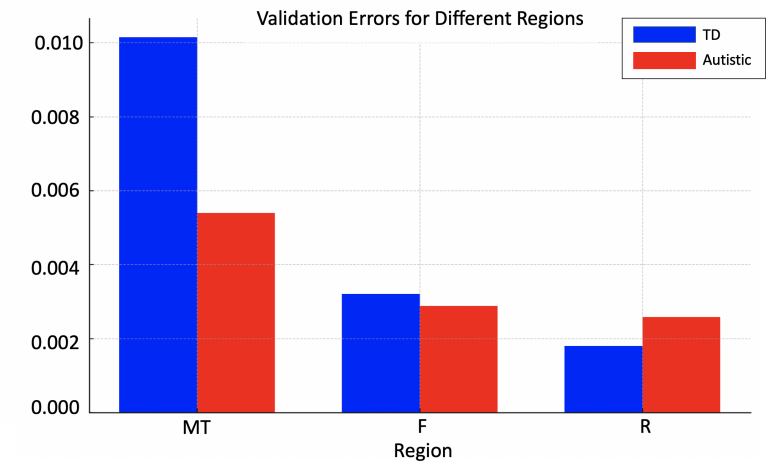

Figure 6: Prediction errors in different regions

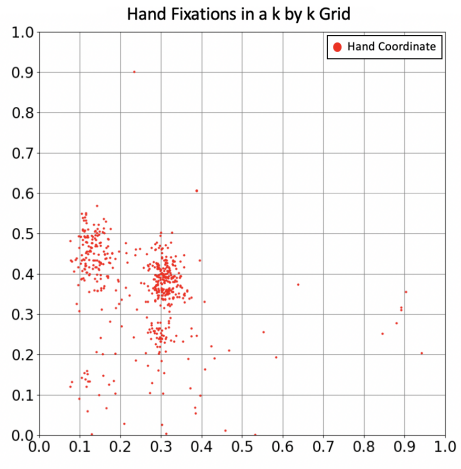

Figure 7: Prediction based on a $k \times k$ grid

Another method we used was splitting the 1×1 grid into k^2 squares, for varying values of k in Figure 7. We labeled the regions from 1 to k^2 and encoded them as k^2 one-hot vectors. Then, we used the start and end times along with the (x, y) coordinates of eye fixations to predict the sequence of regions of hand fixations. We ran our tests with latent dims=150 and $1,000$ epochs with the LSTM, which demonstrated convergence. We also used 80% of our data as training data and used the other 20% as validation/test data. We used the Adam optimizer and the cross entropy loss function.

Table 4: Validation losses from predicting which of the k^2 regions the hand fixations fall in.

k	**10**	**20**	**50**	**100**
Loss (Aut)	1.401	2.083	3.649	5.052
Loss (TD)	2.262	3.068	4.523	6.166
T-statistic	-54.53	-23.55	-35.58	-53.82
P-value	1.419e-11	1.125e-08	4.267e-10	1.575e-11

It can be seen that across all k values, the mean autistic loss is significantly less, with extremely small p values, than the TD loss. This is consistent with our previous findings. The reason for the increasing errors as k increases is due to a larger number of possible outputs of prediction. Although the errors are large, more training data can help improve the prediction accuracy. This method of predicting the general region of hand fixations can be especially useful for intervention, rather than predicting a continuous location exactly, where fixations that lie in specific regions can indicate when attention needs to be redirected.

D.2 Comparison with Alignment-Based Models

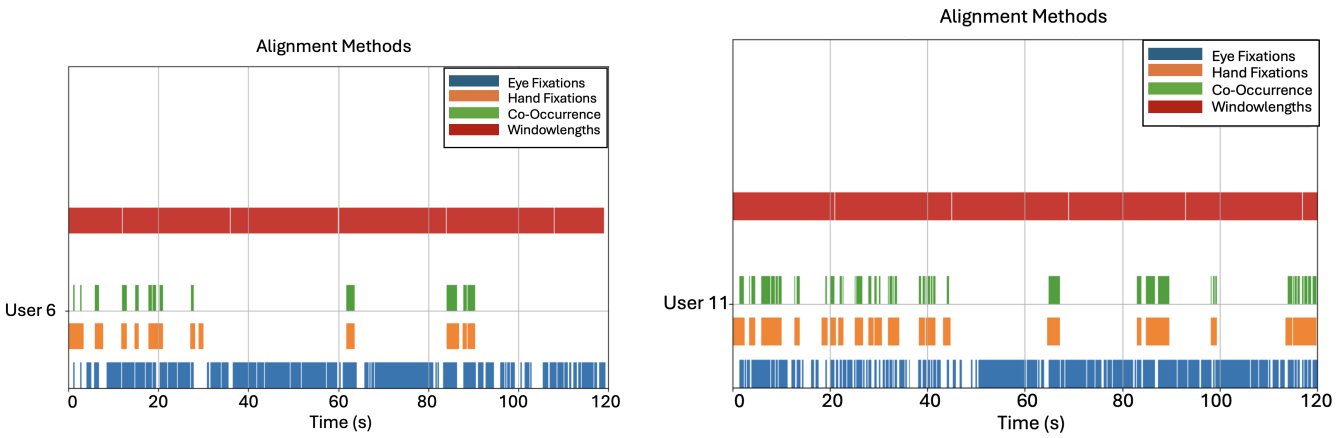

Figure 8: Visual representation of different data alignment: Co-Occurrence (CO) based and Windowlength (WL) based methods (created by student researcher)

A significant source of noise in the data is the realignment of hand and eye fixations: it is difficult to exactly map an eye fixation to a hand fixation. Thus, by utilizing the S2S model, a significant advantage gained is that data alignment is unnecessary. To demonstrate this, I design a data alignment method as a baseline, which can be compared with the proposed model. I propose two alignment techniques, including window length-based and co-occurrence-based.

1. The **Windowlengths (WL)** method aligns the data by splitting up the sequence into time windows and looking at the hand and eye fixations during the specific time windows. In each time window, I assign a specific eye and hand position based on the fixations occurring in that interval, giving us eye and hand sequences aligned by time with the same length. This method made the data a bit less noisy by making the input and output sequences the same length, and each window would correspond from an eye fixation to a hand fixation.

2. The **Co-Occurrence (CO)** based alignment method takes coordinates in the input where there is both an eye and hand fixation to be used for prediction. This method also removes some of the nosiness of the data, making both sequences the same length and providing a more direct matching for the translation. These methods are visualized in Figure 8.

However, these two processes of data alignment both have weaknesses. The Windowlengths method loses some important data values since each window only corresponds to one fixation. Especially in more concentrated data such as in User 11 as shown in Figure 8 (right), assigning one coordinate to each time window loses a lot of information. In the Co-Occurrence method, a lot of data is also lost, especially in sequences where few hand and eye fixations overlap. For example, in User 6 as shown in Figure 3 (left), the data is very sparse, and there are few time intervals with both a hand and eye fixation.

The results are shown in Table 5. We compared the prediction errors between the autistic and typical development groups, and calculated the T-statistic and P-Value to identify the significant differences between the two groups. We see that due to the nature of the sequence-to-sequence model, our method of directly inserting the eye and hand coordinates produced significantly less error across both groups in both models. When taking only the aligned data, or coordinates where there is hand and eye fixation co-occurrence we had a larger error. This is likely due to the large amounts of information lost, especially in sparse data. Additionally, hand fixations even when there isn't an eye fixation and vice versa may be useful predictors of future fixations, which were not included in the input and output sequences. The Windowlengths method had the largest errors, across every time interval, which is likely due to the information lost by assigning a time interval with only one coordinate, especially problematic when the fixations are more concentrated. We see that these LSTM/GRU models are able to effectively handle unaligned data, even performing significantly better than when given pre-aligned data as shown.

Table 5: Model Results (E/H: no alignment)

Error Value	LSTM WL	LSTM CO	**LSTM E/H**	GRU WL	GRU CO	**GRU E/H**
Aut Error	0.0172	0.0104	**0.006351**	0.0199	0.0122	**0.006265**
TD Error	0.0210	0.0268	**0.01204**	0.0208	0.0274	**0.01113**
T-statistic	-11.550	-131.568	**-47.64**	-2.9527	-42.653	**-34.48**
P-Value	1.155e-13	6.746e-50	**1.161-6**	0.005	2.027e-32	**4.217e-6**