Facilitating the Looking Time Paradigm with Machine Learning Models: What to Look Beyond Looking Time?

Sharon Lee

https://github.com/sharonal10/LT

Abstract

The looking time paradigm plays a fundamental role in developmental psychology as a means to test hypotheses about the underlying cognitive process of infants or children. Meanwhile, other types of responses from the subjects, such as facial expressions, body movements, etc., could reveal additional information about their cognitive states. In this work, we investigate whether these features can form a basis for testing hypotheses in developmental science by analyzing raw videos recording infants' behaviors. These videos are much higher dimensional than the looking-time measurement and typically cannot be consumed by standard statistical tools for hypothesis testing. To address this challenge, we develop a hypothesis-testing framework using machine learning models that automatically process features from videos. We evaluate this framework using video data from prior established studies. We first validate the proposed framework in prior studies, by showing that similar conclusions can be reached by using our paradigm and the looking time paradigm. Then, we apply our paradigm to scenarios where the looking time information is unavailable. Further analyses show that our models learn to extract certain spatial and temporal features from videos to make predictions. This suggests that features beyond looking time that can distinguish subjects' behaviors under different experimental conditions exist. Results suggest that machine learning models can effectively extract signals from videos of infant behaviors and can be used for hypothesis testing in developmental psychology studies.

1 Introduction

A core research challenge in developmental research is understanding the mental process of infants and young children underlying the formation of their behaviors. The mental process is unobserved, and hypotheses regarding the process are typically tested with experimental measurements designed by the experimenters. For instance, in a typical looking time experiment, as illustrated in Figure 1, a subject is presented with a visual stimulus designed for an experimental condition, and measurements e.g. looking time are recorded. The effects of experimental conditions are estimated based on the measurements.

The looking time paradigm has its limitation when applied to studies that aim at understanding the cognitive processes in natural tasks, environments, or stimuli, which urges us to seek alternatives. Early developmental studies adopted controlled stimuli in experimental designs [3, 2, 24, 21, 17], where infant reactions to the stimuli are measured by a scalar measurement of looking time. However, in natural tasks and environments, more complex, less controlled stimuli are often inevitable. The subject's responses to these stimuli are likely more elaborate and cannot be easily captured by a few scalar measurements. Other sources of information, such as facial expressions [22], could potentially contain signals reflecting the cognitive process of interest. Here we advocate for using raw observations in the format of video recordings of subjects for hypothesis testing, since they are more complete measures of subjects' behaviors compared to scalar measurements.

This approach enables exciting possibilities but also brings a significant challenge: performing statistical inference on these high-dimensional data for hypothesis testing. Luckily, recent advancements in machine learning has shown a viable path. Deep learning models, i.e., large neural networks, are capable of learning to perform prediction tasks from data [15]. The effectiveness of these models in processing high-dimensional data has been shown in visual recognition

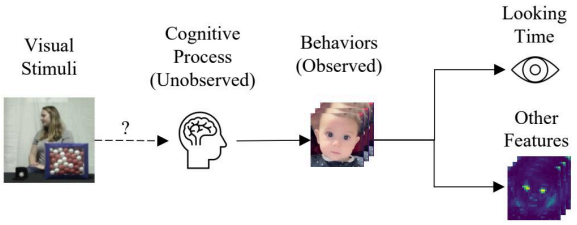

Figure 1: In an experimental trial studying infant subjects, a visual stimulus is shown to a subject whose behavioral reactions are recorded. Our goal is to understand whether the stimulus impacts the subject's unobserved cognitive process. In the preferential looking paradigm, looking time is coded from observed behaviors in the form of video recordings and is used for hypothesis testing. In comparison, we propose to use features directly extracted from video data with the tool of machine learning models.

	Preferential Looking	Logistic Regression	Binary Classification with NNs
X	Looking time	Looking time	NN features extracted from video recordings
Y	Type of visual stimuli		
Null Hypothesis H_0	$\mathbb{E}[X \mid Y=0] = \mathbb{E}[X \mid Y=1]$, i.e. Y does not affect X	$\theta = \theta_0$ (constant model), i.e. X does not affect Y	$H_0 : \mathbb{E}_X\left[\hat{Y} = Y \mid X\right] = 1/2$, i.e. the model has chance-accuracy
Decision Rule	Paired t-test	$g(X;\theta) :\approx p(Y=1 \mid X)$; Likelihood ratio test / Wald test	$g(X;\theta) :\approx p(Y=1 \mid X)$; Binomial test

Table 1: Comparison of statistical inference tools for hypothesis testing.

tasks, in which they have achieved human-level performance in recognition tasks such as image classification [7] and action recognition [8]. Perhaps a more exciting opportunity is that deep learning models can be used to enhance scientific discovery and hypothesis testing [19], as shown by studies in biology [11], neuroscience [14], and material science.

Machine learning approaches have not yet been extensively applied to developmental psychology due to two challenges. The first is non-unified stimulus across different studies. The second is the small size of the data. We addressed these challenges by proposing a hypothesis testing framework that is agnostic to the specific stimuli presented in the experimental designs in developmental psychology studies. Specifically, we propose a pipeline to train deep neural networks to extract information from video recordings of test subjects, together with a statistical inference procedure that uses the trained deep neural networks to test the hypotheses in existing studies. Notably, the formulation of the proposed framework only relies on experimental conditions as labels and does not require any extra data annotations.

We show that the proposed framework effectively utilizes raw video data and provides a low-cost approach to reproduce findings from prior studies, presenting a high correlation with the effect sizes reported using classical looking time analysis methods. Further neural network feature interpretation results with the proposed method suggests that networks attend to particular regions of faces, such as cheeks, when making predictions.

2 Method

2.1 Hypothesis Testing

A looking time study typically involves a fixed set of visual stimuli as experimental conditions. In each experimental trial, a subject receives one type of stimulus and their reaction is observed and recorded. We denote the random variable for their reaction as X, and the types of conditions as Y. X is high-dimensional and is typically challenging to be analyzed directly. Let $f(X)$ be features extracted from raw observations X.

We are interested in testing the following hypothesis:

$$H_0 : X \perp Y. \tag{1}$$

In this section, we will derive three statistical inference procedure, each testing a null hypothesis, rejecting which is sufficient to reject H_0, i.e. all tests are conservative. We will start from the classic approach for looking time analysis,

then introduce a logistic regression model that produces test statistics based on looking time as an alternative, and finally scale up the inference procedure to directly consume video data instead of exclusively relying on looking time as the only feature. Table 1 contains a summary of these approaches.

Mean Group Differences of Looking Time. In the following discussions, we restrict $Y \in \{0, 1\}$ to be a binary class, but the formulation can be easily extended to more conditions.

As a classical approach in looking time studies, f is instantiated as looking time coding, and the following hypothesis is tested:

$$H_0^{\text{LT}} : \mathbb{E}_X \left[f(X) \mid Y = 0 \right] = \mathbb{E}_X \left[f(X) \mid Y = 1 \right]. \tag{2}$$

Logistic Regression Models with Looking Time Inputs. H_0^{LT} aims to test the statistical independence of $f(X)$ and Y. Alternatively, to test the independence, one could test how much $f(X)$ explains Y by fitting the following logistic regression model:

$$`p_\theta(\hat{Y} = 1 \mid f(X)) = S(\theta \cdot [1 \quad f(X)]^T), \tag{3}$$

where $S(u) = 1/(1 + e^{-u})$ is the Sigmoid function, and $\theta \in \mathbb{R}^2$ is the parameter of the regression model. Statistical tests, e.g. a likelihood ratio test, test against the following null hypothesis:

$$H_0^{\text{LR}} : \mathbb{E}_\Theta \Theta = \theta_0, \tag{4}$$

where Θ is the random variable for θ, $p(Y \mid f(X), \Theta = \theta) = p_\theta(Y \mid f(X))$ (Equation (3)), and $\theta_0 = [\mathbb{E}Y \quad 0]$ is the constant predictor.

Deep Neural Networks with Video Inputs. Instead of restricting f to looking-time coding, we explore an alternative paradigm design to explore f in the form of a neural network to process X in the form of video recordings:

$$p_\theta(\hat{Y} = 1 \mid X) = S(\theta \cdot [1 \quad f(X; \theta)]^T) := g(X; \theta), \tag{5}$$

where f is the feature representation extracted by a neural network parameterized by θ. Note that a naïve adaptation of logistic regression models, with $f(X) = X$, is less applicable due to the high dimensionality of X.

Therefore, we aim to test the following hypothesis:

$$H_0^{\text{NN}} : \mathbb{E}_{X,Y} \left[p_\theta(Y \mid X) \right] = \mathbb{E}_{X,Y} \left[p_{\theta_0}(Y) \right], \tag{6}$$

where θ_0 is the optimal constant predictor.

In practice, to evaluate Equation (6), we compute the following:

$$\mathbb{E}_{X,Y} \left[p_\theta(Y \mid X) \right] = \mathbb{E}_{X,Y} \mathbb{E}_{\hat{Y}} \left[\hat{Y} = Y \mid X, Y \right] =: \text{Acc}(\theta), \tag{7}$$

where $\hat{Y} := \arg\max p_\theta(\hat{Y} \mid X)$.

Equation (6) can be re-written as:

$$H_0^{\text{NN}} : \text{Acc}(\theta) = \text{Acc}(\theta_0). \tag{8}$$

Here, H_0^{NN} is equivalent to the hypothesis that given any neural network parameter θ, the prediction accuracy of a trained neural network is equal to that of the optimal constant predictor. For binary conditions, $\text{Acc}(\theta_0) = 1/2$. Rejecting H_0^{NN} is sufficient to reject H_0.

Equation (5) is in analogy to Equation (3), but the high dimensionality of parameters θ prohibits the use of statistical tests as used in Equation (4). Therefore, we test the following hypothesis:

$$H_0^{\text{NN}} : \mathbb{E}_\Theta \text{Acc}(\Theta) = \mathbb{E}_\Theta \text{Acc}(\Theta_{\text{init}}). \tag{9}$$

In Equation (9), the distribution of Θ is estimated via data resampling.

3 Results

3.1 Experiment 1: Reproducing Prior Looking Time Studies

We first conduct a confirmatory analysis to show that the proposed hypothesis testing framework can reproduce results from previous studies. In the following experiments, we apply the proposed framework to test the null hypothesis that subjects cannot distinguish the conditions present in a study.

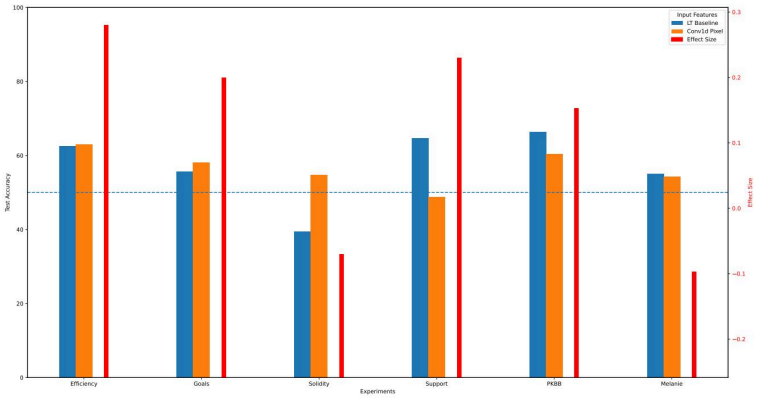

Figure 2: Comparison of test accuracy (50% is chance) with effect sizes reported in original studies (0.0 is no effect). Results for two variants of the proposed framework are reported, with pixel inputs and looking time inputs. Both present a high correlation with the effect sizes from looking time analysis.

Datasets. In this section, we use datasets from prior literature and use the proposed framework for hypothesis testing. We use the following studies: [23] which provides the four datasets ("Efficiency", "Goals", "Solidity", and 'Support"), Gal et al., 2023 [20] ("PKBB"), and [4] ("Melanie").

Results. To compare with the effect sizes reported in prior studies, we compute the correlation between the test accuracy from Equation (9) with the effect sizes. We compare results of the proposed framework under two variants, where f from Equation (5) is instantiated as 1) a deep neural network with video inputs (f_{pixel}), and 2) as a looking time extractor followed by one logistic regression layer (f_{LT}). Details of the inputs are deferred to Section 5. Note that the second variant does not have access to the raw video information and utilizes the looking time information only.

Results are shown in Figure 2. We observe that the pixel variant tends to have a higher test accuracy than the looking time variant when the effect size from looking time analysis is low, which can be attributed to the signals from prominent head and body movements for the former variant that are absent for the latter.

3.2 Experiment 2: Disentangling Signals from Looking Time and Other Video Features

In this section, we aim to disentangle the signals from looking-time-related features and from other features extracted from input videos. We make the following hypotheses for the f_{pixel} model variant from Section 3.1:

1. Video inputs contain information on looking time which is implicitly used by the model for predictions. In particular, the model extracts the looking time information mostly from video lengths, as there is a high correlation of video lengths with looking time (Figure 3).
2. Removing the looking time information would decrease prediction accuracy.
3. Providing other looking-time-related features, such as per-frame annotations of whether the subject is looking at the screen and of looking directions, leads to more explicit signals than raw videos that improve prediction accuracy.

We conduct the following experiments to verify these hypotheses.

Removing Video Length Information. As shown in Fig. 3, there is a strong correlation between total looking time and video length. Since all videos are padded to match the longest video length, they are never shortened. Consequently, the total looking time information can implicitly be used by the model for classification.

In our previous setups, videos are often padded to match the length of the longest video in the dataset which contains looking time information implicitly. In contrast to that approach, we instead employ mean pooling to normalize videos into three consistent lengths, denoted as L_1, L_2, and L_3. After the convolutional layers, a mean pooling layer is applied on sliding windows across the temporal dimension of the video frame features. Given a sequence of feature maps $F = \{f_1, f_2, \ldots, f_N\}$ extracted from N frames, mean pooling over a window of length x at position i (where $i + x \leq N$) is calculated as $MP_i = \frac{1}{x} \sum_{j=i}^{i+x-1} f_j$. Fig. 4

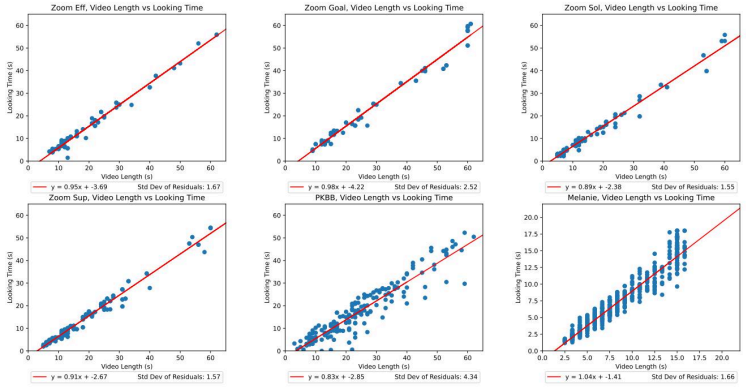

Figure 3: The x-axis represents the video length, while the y-axis indicates the total looking time for each video in the dataset.

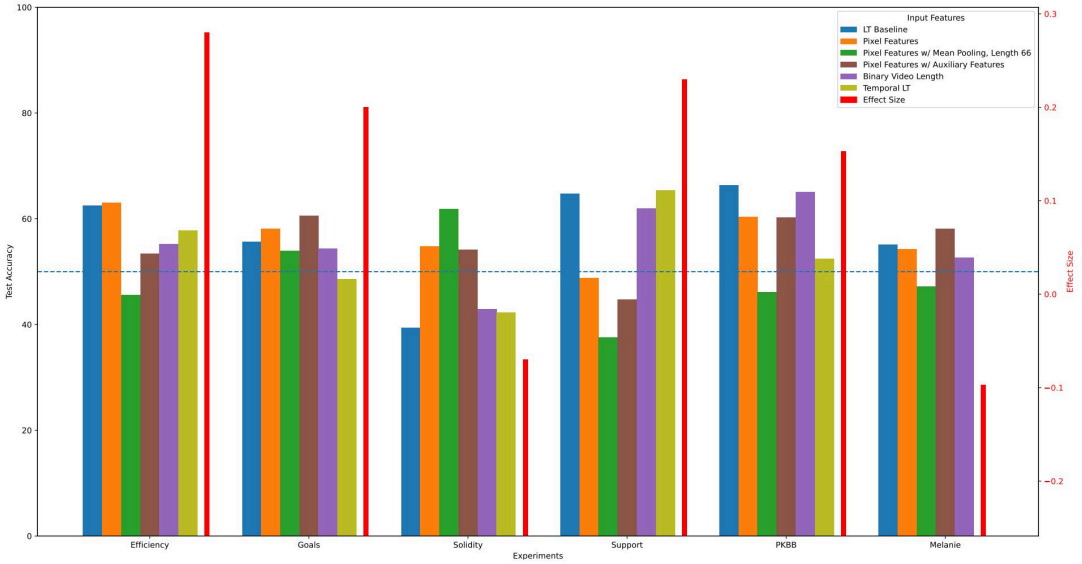

Figure 4: X-axis represents the dataset, while the y-axis indicates the test accuracy and magnitude of the effect size.

Results show that mean pooled pixel features with CNN model results are strictly worse than our pixel features with CNN model with the exception of the solidity dataset where there is a low effect size between total looking time, suggesting that the video length provides learning signals for model prediction.

Importance of temporal looking time Additionally, as total looking time is an important feature in previous studies, we examine the importance of temporal looking time. More particularly, instead of only using the total amount of time per experiment, we use annotations for every frame where 0 is when the infant is looking away and 1 is when the infant is looking at the screen. Intuitively, this would mean that an infant looking at the screen for n seconds glancing back at the screen and away every 1 second, would be differentiated from an infant who is looking at the screen for n consecutive seconds. We show that temporal looking time may be useful as it increases the accuracy in the Pokebaby dataset where the temporal looking time is manually annotated.

Importance of Looking Directions. Moreover, we explore the importance of looking directions by examining only the importance of looking directions that are manually annotated in the Pokebaby dataset (Table 5). Here, we hypothesize that looking directions does help the overall classification accuracy.

Eff	Goal	Sol	Supp	PKBB	Melanie
100%	100%	93.4%	94.7%	99.4%	98.1%

Table 2: Percentage of videos where the highest activation spike for Conv3d occurred at the very end of the video.

Overall, experiments above show that 1) there is a strong correlation between total video length and looking time, 2) removing the implicit looking time information decreases prediction accuracy, indicating that looking time information is crucial for final classification, and 3) explicit temporal signals from other looking-time related features which indicate when the subject is looking at the screen versus looking away improve the final classification accuracy.

3.3 Experiment 3: Feature Analysis via Activation Maps

In order to make the correct prediction, the model is trained to extract representative features from videos that are informative for the prediction task, and we investigate whether the discovered features are independent of or correlated with looking time. The new formulation can also be considered an automatic feature selection paradigm, which permits a further investigation of learned features $f(X) = f(X;\theta)$ where θ is the parameterization of a trained neural network. Note that our model is trained on data from a set of subjects but tested on a novel set of subjects. This enables the model to generalize to novel instances. What is being learned by the model, is not features specific to the subjects it is trained on, but rather generalizable patterns or features. We apply interpretation techniques for $f(X)$.

Temporal Activations. As briefly mentioned in Section 3.1, in Figure 3, there is a strong correlation between total looking time and video length. As the experiment terminates after the infant stops looking at the screen for 2 seconds [20], we see that video length information correlates to total looking time. We show Experiment 1 and Experiment 2 that the model achieves low performance without video length. We hypothesize that the model uses video length information for the final classification.

Here, we explore activations across frames to address: (1) if the activation spikes are highest at the boundary of video length Table 2 (2) if the average activations is higher in the padding frames compared to the non-padding frames; here our hypothesis is that if the model uses any pixel information from the infant video, the average activation of the non-padding frames should be higher than the average activation of the padding frames. Here, we exclude the boundary between the non-padding and padding frames.

For most datasets, the highest activation spike for Conv3d occurred at the video length i.e. the boundary between the padded and non-padded frames Table 3. This suggests that for these datasets, video length plays an important part in the predicted class. We also see that the non-padding activations are always higher than all padding frames which shows that the model uses pixel information for its final classification.

	Eff	Goal	Sol	Supp	PKBB	Melanie
Non-padding	619.7 ± 235.1	458.5 ± 194.4	1972.8 ± 933.8	1388.5 ± 360.0	736.3 ± 232.9	3067.4 ± 682.9
Padding	360.7 ± 0.7	304.7 ± 0.7	1256.4 ± 1.3	855.7 ± 1.0	783.6 ± 0.9	1538.7 ± 0.5

Table 3: The average activation of non-padding and padding frames for Conv1d.

The average activation of padding frames for almost all datasets is higher than the average activation of non-padding frames. This suggests that despite video length playing a large part in the predicted class, the features within the non-padded frames is also important.

Within the padded area of the videos, we have also compiled a dataset that marks frames falling in the top 5%, 10%, and 15% activation percentiles. From our analysis, we see that these frames indicate large movements/ rotations in the babies face, or in their expressions.

Spatial Activations. As we use pixel features to train a CNN for classification, we further look into regions of each 2D frame to identify highly activated regions of the infant's face. To do this, we use DINO [5, 18] as pretrained cosegmentation vision model [1], to segment prominent facial parts consistently across all videos per dataset. Each segment was converted into a binary map and multiplied by the activation values of each pixel within that segment. The segment labels were manually labeled for semantics. For each segment, we aggregated the average activation values by calculating the mean activation for all pixels within the segment. Results are shown in Table 4.

Normalized Average Activation	Eff	Goal	Sol	Supp	PKBB	Melanie
Eyes, Mouth, Ears	0.77	0.80	0.95	0.94	0.94	1
Cheeks	1	1	1	1	0.93	0.15
Forehead	0.71	0.74	0.91	0.93	0.83	0.65
Hair	0.95	0.92	0.97	0.92	0.83	0.65
Background	0.89	0.80	0.96	0.92	1	0.47

Table 4: Normalized Average Activation for DINO Segmented Regions. Average activation scores were normalized per-dataset by setting the score of the cluster with the largest average activation to 1 and scaling the other values accordingly.

Figure 5: DINOv2 segmentation.

We aim to interpret the working procedure of the model for classifying infant videos on a per-frame basis. This interpretation helps verify that the model is focusing on specific regions of the video to discover new features in the infant's face that may be important. We use activation maps to visualize the regions of the face (segmented by a pretrained segmentation model, DINOv2) that are most highly activated (Fig. 5). In this specific frame, we show that the cheeks is the most activated region. Within each frame of the video, we investigate the regions of the face that are most highly activated. We explore the activation within the non-padded videos across frames by examining the most highly activated frames per video and the activation of 2D regions within each frame. The aggregated results of the most highly activated regions from each frame are compiled across all frames per video for all videos in the dataset and are shown in Table 4.

Across all datasets, we see that cheeks are most highly activated in 4 out of 5 of the datasets and the forehead is the region which is least highly activated in 4 out of 5 of the datasets. Notably, the Melanie dataset has much lower activation in the cheek region compared to the eyes, mouth, and ears. This may be due to the high contrast between the illuminated front of the face and the shadowed sides of the head for most videos in that dataset.

3.4 Experiment 4: Ablation Study

We also carry out ablation trials on architectures such as the CNN, transformer, and neural network architectures as described in Section 5.3, which use various features such as pixel features, landmark features, temporal looking time and direction, as well as total looking time and direction as described in Section 5.2.

The table in Table 5 presents the key quantitative findings of our experiments on the Pokebaby dataset. For facial landmark features, the one-layer NN achieved the highest performance (65.4 ± 2.4), outperforming both the CNN (61.6 ± 2.0) and Transformer (57.0 ± 2.9). Temporal looking time features demonstrated a significant advantage for the one-layer NN (69.6 ± 0.7) over the CNN (49.5 ± 0.7) and Transformer (50.3 ± 0.2). Although the best performing model in the Pokebaby dataset is the facial landmark features and CNN, we find that this is only true for this dataset due to the high video quality which allows for facial landmarks to be detected more accurately.

	Logistic Regression	SVM
Total Looking Time	66.3	66.3
Total Looking Directions	72.8	74.5
Temporal Looking Time	70.2	71.5

Table 5: Looking Time and Looking Direction Baseline Results for the Pokebaby Dataset

	CNN	Transformer	One-layer NN	Two-layer NN
Pixel Features	60.4 ± 1.1	58.4 ± 5.8	-	-
Facial Landmark Features	61.6 ± 2.0	57.0 ± 2.9	65.4 ± 2.4	63.0 ± 1.8
Temporal Looking Time	49.5 ± 0.7	50.3 ± 0.2	69.6 ± 0.7	68.3 ± 0.4
Temporal Looking Directions	50.8 ± 0.2	49.1 ± 0.8	62.8 ± 1.6	61.6 ± 0.5

Table 6: Comparison of Features and Model Architectures for the Pokebaby Dataset

4 Conclusions

A core problem in developmental research is understanding the mental processes underlying the formation of the preferences or expectations of infants or young children. In this study, we develop a hypothesis testing framework to develop such understanding, relying on the same experimental setup as the looking time paradigm, but basing our analysis on the behavioral features of experimental subjects including but not limited to looking time. We use video data as model inputs throughout the experiments, which are easy to obtain, especially with the increasing popularity of online experiments [6]. The proposed paradigm can also be extended to other data modalities such as heart rate, sucking measurement, EEG signals, etc. as input formats and therefore is not restricted to looking-time studies. We showcase the credibility of the proposed framework by reproducing results from prior looking time studies, establishing a high correlation between the hypothesis testing results from the proposed framework with effect sizes reported using classical looking time approaches. Further analysis on neural network interpretation suggests that particular regions from subject faces, such as cheeks, provide signals for network prediction, and can be viewed as complementary features to looking time.

5 Method

5.1 Training Framework

Model Inputs Inspired by the literature for ranking predictions [13] and the notion of "baseline" in psychological studies [9], we formulate the problem in a machine learning context as follows. As shown in Figure 1, a model is given video recordings of an infant in reaction to some visual stimuli as inputs, and is asked to predict information on the stimuli. Specifically, the model receives a pair of videos from two experimental trials in a looking time study for one infant subject, and is trained to predict if these two videos correspond to the same of different types of stimuli. If the model can make predictions with above-chance accuracy, it suggests that the input videos contain non-trivial information to infer the stimuli, and therefore indicates that infants have the capability of distinguishing among stimuli.

Model Outputs Models are trained to predict the type of stimulus in Figure 1. In every experimental study, two stimuli are always presented to the infant. These stimuli can vary across experiments and might be videos [20], images [10], sounds [12], or an actor demonstrating a behavior [16].

Data Preprocessing. Video data recording infant reactions are first trimmed to only include trial periods, and we follow the criterion of trial termination as specified in original studies. Examples of the criteria include alarm sounds in the video indicating the start of the experiment or lighting changes when videos involving screens begin. Dataset-specific termination criteria are listed in the appendix. Trimmed videos are then cropped with bounding boxes around baby faces. To achieve this objective, we employ a bounding box tracking model, YOLO v8, which tracks all potential bounding boxes of objects of interest within the video. Subsequently, we utilize a language-prompted bounding box extractor, LangSAM, with the prompt 'baby' to select the bounding box associated with the baby. We then calculate the minimum distance between the original bounding boxes generated by YOLO v8 and those identified for the baby by LangSAM. The bounding box with the smallest distance is selected, utilizing YOLO v8's output for continuous tracking of the baby throughout the video. As parent's faces are often included in the video, combining these two methods allow for bounding boxes to only fixate on the baby and not the parent throughout the video. Full bodies are excluded since body movements vary greatly based on the experimental setup (e.g., babies can sit on their parents' laps or on a fixed chair). Moreover, many of the experiments feature camera views that occlude or cut off parts of the babies' bodies, making it challenging to capture consistent body crops. Then, we employ video stabilization using point feature matching to eliminate any unintentional camera movements.

Data Pairing Babies naturally have varied behavior patterns; for instance, some move more than others or might have shorter attention spans leading to a lower average looking time. These behaviors can be influenced by factors in different lab setups or unexpected events during data collection, like parental interference or experimenter errors. Given

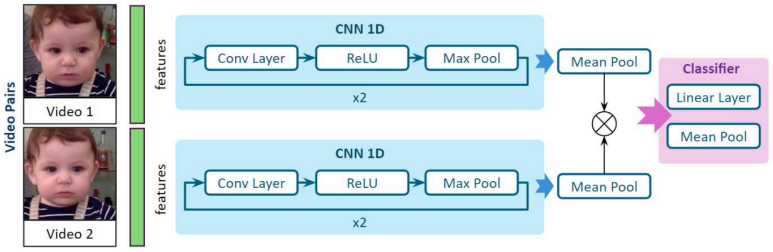

Figure 6: A 1D CNN architecture processes pairs of videos (one negative and one positive) representing the same infant. If one video is a positive sample, the other must be a negative sample; thus, the classifier needs only to predict the label for the first video.

these variables, classifying a video label based solely on its content is challenging. To address this, we pair videos with different labels from each baby and use these pairs as our model's input. Since paired videos have distinct labels, we only give a prediction for the first video in the pair making the second video have the opposite prediction. This means that classifying the first video alone effectively represents the pair's label.

5.2 Features

Our feature ablations are listed below. Each video is processed in order to obtain a set of sequential frame features before performing any training.

Pixel Features After cropping the video to track the infant's face, each frame is normalized and resized to a 32×32 RGB image, resulting in a `num_frames` $\times 32 \times 32 \times 3$ array for each video. Zero-padding is then added to each array to obtain a $1600 \times 32 \times 32 \times 3$ array.

Facial Landmark Features After cropping the video to track the infant's face, the video is passed to an infant facial landmark estimation model which is a HRNet trained on the InfAnFace dataset, consisting of 410 annotated images of infant faces []. This yields a set of 68 coordinates for each frame, resulting in a `num_frames` $\times 136$ array for each video. Zero-padding is then added to each array to obtain a 1600×136 array.

Video Length Features Video Length Features are very similar to Facial Landmark Features. The only difference is that all coordinates are replaced by ones, resulting in a `num_frames` $\times 136$ array of ones for each video. Zero-padding is then added to each array to obtain a 1600×136 array.

Temporal Looking Time Using iCatcher, video processing software which classifies gaze direction in real time, each frame is represented by a binary number representing whether the infant was looking at the trial at that point in time. After padding, this results in a flat array of length 1600.

Temporal Looking Directions Similar to Temporal Looking Time, each frame is represented by a single integer. The values can be 0 (looking away), 1 (left), or 2 (right). We use -1 to represent padding frames as well, as the array is still padded to length 1600. This is only applicable to specific datasets which recorded this information.

Total Looking Time Total Looking Time is also obtained from collected iCatcher data, and represents the total amount of time the infant spent looking at the trial, in seconds.

Total Looking Directions Total Looking Directions is a length 3 array. Each element represents the total amount of time the infant spent looking left, right, or not looking at the trial at all.

5.3 Model Architecture

Our encoder ablations are listed below. Each set of sequential frame features is padded with zeros to a fixed length and input into the encoder, which yields one feature. The fixed length here is the length of the longest video length in the dataset. The two resulting output features are concatenated and then passed through a linear classifier where they are first projected into a lower-dimensional space. This projected feature is then fed into another linear layer to obtain the class logit.

An overview of the architecture can be found in Figure 6.

Transformer Our Transformer model appends a learned label token to the input sequence, processes it using PyTorch's TransformerEncoderLayer, and then extracts the output associated with the label token to use as a representation of the entire sequence.

1D CNN Due to our video pairing paradigm, we have two videos per training/ test sample denoted as Video A and Video B. Video A is passed through a 1d convolutional layer, ReLU activation, and then it is maxpooled and then mean pooled. This is also done for Video B The results are concatenated, then passed through a classifier consisting of a linear layer followed by the Sigmoid function.

3D CNN Our 3D CNN architecture follows a similar pattern - the only difference is that our CNN consists of two sets of Conv3d $\rightarrow$ ReLU $\rightarrow$ MaxPool3d $\rightarrow$ Dropout3d, followed by Linear $\rightarrow$ ReLU $\rightarrow$ Dropout.

One-layer Neural Network Our One-layer Neural Network consists of a single linear layer, followed by a ReLU activation function to introduce non-linearity.

Two-layer Neural Network Our Two-layer Neural Network is very similar to the One-layer Neural Network. It consists of two linear layers, both of which are followed by ReLU activation functions.

5.4 Model Optimization.

As for the training schedule, we use the standard Binary Cross Entropy loss to guide the training for 800 epochs or until the training accuracy is > 0.95 for at least 10 epochs. We decided not to use a validation set during training, as some datasets are small. Across all encoder models, we use the Adam optimization algorithm starting with a learning rate of 1×10^{-6} and implementing a cosine decay schedule to decrease the learning rate to 1×10^{-7}.

5.5 Training Strategy.

In the looking-time paradigm, looking time for each stimulus class is grouped by babies. For each baby, an average looking time is calculated for each stimulus class, and statistical testing is performed to determine if there is a significant difference between the looking times of two stimulus classes. To directly compare our model's performance with the results from the LT Paradigm, we opt for a training scheme called the "Leave One Out" strategy, which also segregates data at the baby level. In this approach, we conduct multiple training and testing runs. In each run, we place exactly one baby into the test set if that baby has not been previously tested, while all other babies are placed in the training set‘ This method allows us to gauge baby-level test accuracy. We then compute the average and standard deviation of the test accuracy across all babies. Lastly, we determine the correlation between the average model accuracy and the statistical significance found in the LT Paradigm.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021.

[2] Renee Baillargeon. Object permanence in $31/2$-and $41/2$-month-old infants. *Developmental psychology*, 23(5):655, 1987.

[3] Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

[4] Krista Byers-Heinlein, Christina Bergmann, Catherine Davies, Michael C Frank, J Kiley Hamlin, Melissa Kline, Jonathan F Kominsky, Jessica E Kosie, Casey Lew-Williams, Liquan Liu, et al. Building a collaborative psychological science: Lessons learned from manybabies 1. *Canadian Psychology/Psychologie Canadienne*, 61(4):349, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. pages 9650–9660, 2021.

[6] Aaron Chuey, Veronica Boyce, Anjie Cao, and Michael C Frank. Conducting developmental research online vs. in-person: A meta-analysis. *PsyArXiv*, 2022.

[7] 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.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[9] Francesco Galvano. Baseline assessment in behaviour analysis: Exploring the differences between general and contextual baselines. 07 2017.

[10] Theresa M. Gerhard, Jody C. Culham, and Gudrun Schwarzer. Distinct visual processing of real objects and pictures of those objects in 7- to 9-month-old infants. *Frontiers in Psychology*, 7, 2016.

[11] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[12] Peter W. Jusczyl, David B. Pisoni, and John Mullennix. Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, 43(3):253–291, 1992.

[13] Lars Kotthoff. Ranking algorithms by performance, 2013.

[14] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[16] Andrew N. Meltzoff and M. Keith Moore. Early imitation within a functional framework: The importance of person identity, movement, and development. *Infant Behavior and Development*, 15(4):479–505, 1992.

[17] Kristine H Onishi and Renée Baillargeon. Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258, 2005.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[19] Maithra Raghu and Eric Schmidt. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*, 2020.

[20] Gal Raz, Anjie Cao, Minh Khong Bui, Rebecca Saxe, and Michael Frank. No evidence for familiarity preferences after limited exposure to visual concepts in preschoolers and infants. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45, 2023.

[21] Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

[22] Conor M Steckler, Zoe Liberman, Julia W Van de Vondervoort, Janine Slevinsky, Doan T Le, and J Kiley Hamlin. Feeling out a link between feeling and infant sociomoral evaluation. *British journal of developmental psychology*, 36(3):482–500, 2018.

[23] Yang Wu and Hyowon Gweon. Surprisingly unsurprising! infants' looking time to improbable events is modulated by others' expressions of surprise, 2019.

[24] Fei Xu and Vashti Garcia. Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, 105(13):5012–5015, 2008.