Beyond Average: Individualized Visual Scanpath Prediction

Xianyu Chen Ming Jiang Qi Zhao
University of Minnesota, United States
{chen6582, mjiang}@umn.edu, qzhao@cs.umn.edu

Abstract

Understanding how attention varies across individuals has significant scientific and societal impacts. However, existing visual scanpath models treat attention uniformly, neglecting individual differences. To bridge this gap, this paper focuses on individualized scanpath prediction (ISP), a new attention modeling task that aims to accurately predict how different individuals shift their attention in diverse visual tasks. It proposes an ISP method featuring three novel technical components: (1) an observer encoder to characterize and integrate an observer's unique attention traits, (2) an observer-centric feature integration approach that holistically combines visual features, task guidance, and observer-specific characteristics, and (3) an adaptive fixation prioritization mechanism that refines scanpath predictions by dynamically prioritizing semantic feature maps based on individual observers' attention traits. These novel components allow scanpath models to effectively address the attention variations across different observers. Our method is generally applicable to different datasets, model architectures, and visual tasks, offering a comprehensive tool for transforming general scanpath models into individualized ones. Comprehensive evaluations using value-based and ranking-based metrics verify the method's effectiveness and generalizability.

1. Introduction

Saccadic eye movements, such as fixations and saccades, enable individuals to shift their attention quickly and redirect their focus to different points in the visual field. Studying various factors driving people's eye movements is important for understanding human attention and developing human-like attention systems. Computational models predicting eye movements have broad impacts across various domains, such as assessing image and video quality [8, 27, 47], developing intuitive human-computer interaction systems [33, 40, 55, 64, 67], creating immersive virtual reality experiences [1, 57, 58], enhancing the safety and efficiency of autonomous vehicles [28, 77, 78], and diagnosing neurodevelopmental conditions [11, 22, 39].

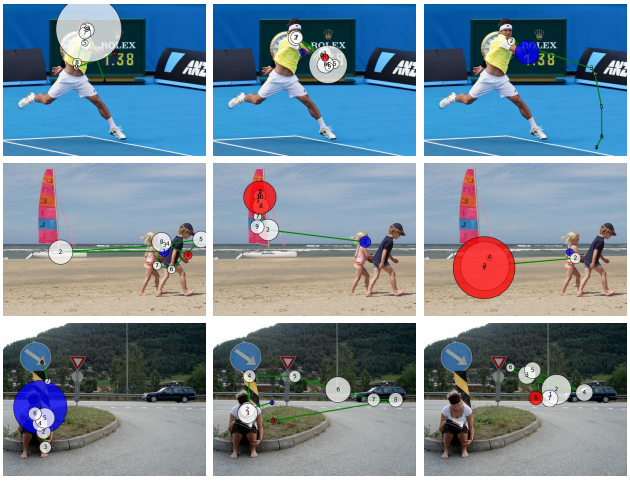

Figure 1. Understanding and predicting the distinct eye movements of each observer is the key objective of individualized scanpath prediction. These examples reveal the variations in the scanpaths of different observers, showing their distinct attention preferences in (a) faces, (b) objects, and (c) background. Each dot represents a fixation, with the number and radius indicating its order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

While existing models of saccadic eye movements predominantly focus on modeling generic gaze patterns manifested as observer-agnostic scanpaths (*i.e.*, a spatiotemporal sequence of fixations), this work seeks to model the individual variations in eye movements. As shown in Figure 1, there exists significant inter-observer variations in visual scanpaths. Such variations can be attributed to a multitude of individual characteristics, such as gender, age, and neurodevelopmental conditions [56, 61]. For instance, females show more explorative gaze patterns than males [53, 62, 63], older adults prefer faces [54] and objects with high color visibility [74], individuals with neurodevelopmental disorders, such as autism spectrum disorder (ASD), may show a preference for repetitive patterns

while avoiding social cues [45, 66, 71]. Therefore, developing tailored models that cater to the uniqueness of each observer is an essential step toward more precise and adaptive attention modeling.

Existing research efforts have failed to address the divergence between the personalized nature of human attention and the collective nature of current scanpath models. This is due to the lack of standardized methods for quantifying and representing individual attention traits, as well as the absence of comprehensive frameworks that can accommodate the diverse range of observer characteristics. In this paper, we resolve this significant challenge with a novel individualized scanpath prediction (ISP) method comprising three novel components: (1) The observer encoder is a key component for personalized scanpath modeling. It efficiently captures an observer's unique attention traits by introducing an observer-specific identifier as an additional input, forming the basis for individualized scanpath predictions. (2) The observer-centric feature integration module adopts a comprehensive approach, fusing visual features, task guidance, and observer-specific attention traits spatially and channel-wise. This ensures consideration of diverse bottom-up and top-down cues, simplifying subsequent processing and enhancing the efficient prediction of individualized scanpaths. (3) The adaptive fixation prioritization module enhances scanpath precision by dynamically assigning priorities to the output features, generating a probability map for each fixation. This adaptability ensures refined predictions of individualized scanpaths.

Our method has three distinctions from previous visual scanpath studies: (1) We go beyond prior work focusing on general scanpath modeling and propose the first comprehensive investigation of individualized scanpath prediction. (2) We emphasize the tight integration of observer features into the scanpath prediction process, distinct from trivial individualization techniques such as fine-tuning with single-observer data. (3) Our method is generally applicable to various model architectures and visual tasks, broadening its usability in real-world applications.

The main contributions of this work are as follows:

1. We study the underexplored task of individualized scanpath prediction, focusing on modeling how an observer's unique attention traits affect their eye movements.
2. We propose an individualization method featuring three novel technical components: The observer encoder is an important addition to scanpath models, which enables observer-centric feature integration and adaptive fixation prioritization. These components enable the model to adapt to individual observers, yielding accurate and individualized predictions.
3. We comprehensively evaluate scanpaths from individual observers' perspectives, using both value-based and ranking-based metrics. Experimental results on multiple eye-tracking datasets, with different model architectures and visual tasks, prove our method's effectiveness and generalizability for predicting individualized scanpaths.

2. Related Works

Our work is related to prior studies on eye-tracking datasets and visual scanpath prediction methods.

2.1. Eye-Tracking Datasets

The foundation for attention modeling relies on diverse, thoughtfully curated eye-tracking datasets spanning various stimuli, tasks, and observers [12, 22, 71, 79, 83]. These datasets, from those dedicated to free-viewing [22, 71, 79] to those capturing goal-directed behaviors [12, 83], serve as invaluable resources for training and evaluating attention models. Specifically, several well-recognized eye-tracking datasets have provided benchmarks to quantify the performance of saliency models [6, 37, 41, 79] and scanpath models [22, 71, 79]. Subsequent studies have developed datasets of goal-directed behaviors to characterize how observers search for an object in an image [83] or answer image-related questions [12]. These efforts facilitate the development of static saliency models [4, 9, 13, 15, 25, 31, 35, 43] as well as dynamic scanpath models [14, 19, 51, 59, 68, 69, 83–85]. Our work sets itself apart from individualized saliency models [13, 46, 49, 52, 80, 81] by predicting dynamic scanpaths rather than static saliency maps. It utilizes datasets from various visual tasks and observer groups to expand scanpath modeling, with emphasis on the distinct attention traits of each observer.

2.2. Visual Scanpath Prediction

Scanpath prediction has been an underexplored topic in the field of attention modeling. Early studies generate scanpaths by sampling fixations from saliency maps using the inhibition-of-return mechanism [34, 46, 50, 72, 73, 75]. Recent studies have developed computational models directly predicting the sequence of fixations and saccades [14, 19, 51, 59, 69, 83–85]. Several scanpath models harness the power of deep neural networks [14, 19, 40, 44, 51, 59, 69, 83–85], reinforcement learning techniques [14, 83, 84], and transformer-based models [51, 59], ultimately improving the accuracy of scanpath prediction to the human level. These developments have significantly deepened our understanding of the temporal dynamics of human attention. However, existing models focus on predicting general scanpaths rather than taking individual variations into account. Differently, our method places particular emphasis on characterizing individual attention traits and integrating them into a general scanpath model, thus enabling tailored predictions that align with each observer's gaze behavior. This unique approach extends the horizon of attention modeling,

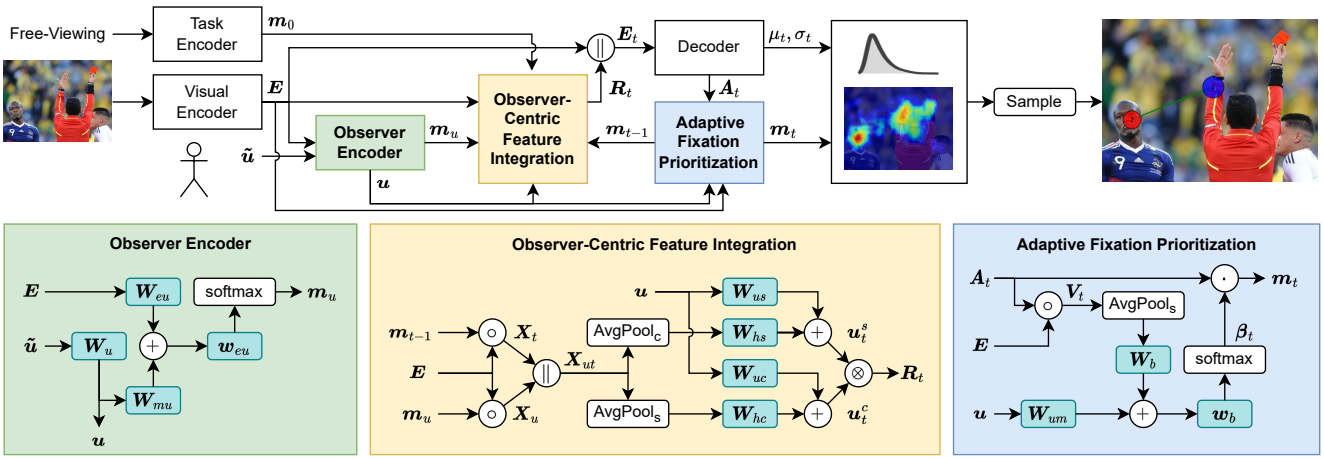

Figure 2. Our proposed method incorporates an observer encoder for characterizing individualized attention traits, followed by observer-centric feature integration for holistic processing, and adaptive fixation prioritization for refined predictions.

underlining the importance of individual differences within the broader context of human attention.

3. Methodology

The core challenge in individualized scanpath modeling is the need to predict unique gaze patterns for different observers. This arises due to the inherent variations in attention traits. Figure 2 presents an overview of our method. It offers a threefold solution: (1) an observer encoder, (2) an observer-centric feature integration module, and (3) an adaptive fixation prioritization module. These components are designed to be flexible as they can be applied on a general scanpath model based on the encoder-decoders, (*e.g.*, with a visual encoder and task encoder, and an LSTM [30] or Transformer [20] decoder) to provide robust and precise predictions tailored to each observer.

3.1. Observer Encoding

At the core of our proposed method is an *Observer Encoder*, a key component designed to enable the novel task of individualized scanpath prediction. It takes as input an observer-specific identifier $\tilde{\boldsymbol{u}}$ (*e.g.*, a one-hot vector) and efficiently computes an observer feature $\boldsymbol{u}$. This feature represents the unique characteristics and preferences of each observer. Our approach utilizes a linear embedding operation to derive the observer feature:

$$\boldsymbol{u} = \boldsymbol{W}_u \tilde{\boldsymbol{u}}, \tag{1}$$

where $\boldsymbol{W}_u$ indicates learnable parameters. The linear embedding operation provides a straightforward mapping that retains important characteristics, offering a practical and computationally efficient solution for capturing unique attention traits.

This observer encoder can be seamlessly integrated into an existing scanpath model. As shown in Figure 2, a typical *Visual Encoder* is used to transform the input image into multi-channel feature maps $\boldsymbol{E}$ characterizing the bottom-up attention. To model the interaction between the visual feature $\boldsymbol{E}$ and the observer feature $\boldsymbol{u}$, an observer guidance map can be computed through a linear combination:

$$\boldsymbol{m}_u = \mathrm{softmax}\big(\boldsymbol{w}_{eu}^T \tanh(\boldsymbol{W}_{eu}\boldsymbol{E} + \boldsymbol{W}_{mu}\boldsymbol{u})\big), \tag{2}$$

where $\boldsymbol{w}_{eu}$, $\boldsymbol{W}_{eu}$, $\boldsymbol{W}_{mu}$ are learnable parameters. This observer guidance localizes salient image regions of specific interest to the observer.

Some scanpath models use a *Task Encoder* to process task-relevant information guiding the gaze behavior, such as a search target or a general question to answer. Such top-down guidance can be represented as a spatial attention map $\boldsymbol{m}_0$ prioritizing task-relevant regions. These bottom-up and top-down features are typically processed with a decoder (*e.g.*, LSTM or Transformer) to predict a sequence of probability maps $\{\boldsymbol{m}_t \,|\, t = 1, 2, \ldots, T\}$ and distribution parameters $\{(\mu_t, \sigma_t^2) \,|\, t = 1, 2, \ldots, T\}$ for sampling fixation positions and durations, respectively, where T is the number of fixations.

In Sections 3.2 and 3.3, we present specific modules that leverage the observer feature $\boldsymbol{u}$ to individualize the scanpath model. For clarity, our method description focuses on its integration with an LSTM model [14]. Please refer to Section 4.1 and Supplementary Material for details about its adaptation to a Transformer network [51].

3.2. Observer-Centric Feature Integration

With the encoded observer features characterizing each observer's distinct attention traits, we design observer-centric feature integration to address the critical need to fuse various inputs, including visual features, task relevance, and

observer-specific characteristics, into a unified representation. The motivation behind this integration is to create a comprehensive understanding of individualized attention patterns. This integration process results in a sequence of observer-centric feature maps $\{\boldsymbol{R}_t \,|\, t = 1, 2, \ldots, T\}$ representing spatiotemporal fixation patterns, thus enabling the model to track individualized attention dynamics over time [12, 14, 38].

Specifically, to guide the prediction at each step, we leverage the predicted fixation distribution from the previous step (*i.e.*, $\boldsymbol{m}_{t-1}$) as a soft attention map, applying it to the visual features to derive the previously fixated visual features $\boldsymbol{X}_t = \boldsymbol{E} \circ \boldsymbol{m}_{t-1}$, where the symbol $\circ$ denotes the Hadamard product. It is noteworthy that the task guidance map $\boldsymbol{m}_0$ is used initially to guide the first fixation, mimicking the cognitive process that initially directs eye movements based on the visual task. Similarly, the observer guidance map $\boldsymbol{m}_u$ is used as the attention weights to obtain observer-centric visual features $\boldsymbol{X}_u = \boldsymbol{E} \circ \boldsymbol{m}_u$.

To seamlessly integrate the fixated visual features and observer-centric visual features, we concatenate the two types of feature maps

$$\boldsymbol{X}_{ut} = \boldsymbol{X}_t \parallel \boldsymbol{X}_u, \tag{3}$$

and perform spatial and channel-wise feature fusion, which are achieved by average-pooling the feature maps along the channel (AvgPool_c) and spatial dimensions (AvgPool_s), respectively, followed by linear layer processing and the addition of encoded observer features:

$$\boldsymbol{u}_t^s = \text{ReLU}(\boldsymbol{W}_{hs}\text{AvgPool}_c(\boldsymbol{X}_{ut}) + \boldsymbol{b}_{hs}) + \boldsymbol{W}_{us}\boldsymbol{u}, \tag{4}$$

$$\boldsymbol{u}_t^c = \text{ReLU}(\boldsymbol{W}_{hc}\text{AvgPool}_s(\boldsymbol{X}_{ut}) + \boldsymbol{b}_{hc}) + \boldsymbol{W}_{uc}\boldsymbol{u}, \tag{5}$$

where $\boldsymbol{W}_{hs}$, $\boldsymbol{W}_{hc}$, $\boldsymbol{W}_{us}$, $\boldsymbol{W}_{uc}$, $\boldsymbol{b}_{hs}$, and $\boldsymbol{b}_{hc}$ are learnable parameters. Ultimately, combining $\boldsymbol{u}_t^s$ and $\boldsymbol{u}_t^c$ yields the final observer-centric feature maps

$$\boldsymbol{R}_t = \boldsymbol{u}_t^s \otimes \boldsymbol{u}_t^c, \tag{6}$$

where $\otimes$ is the outer product. It represents the dynamic importance of individual attention traits in the prediction of the current fixation, providing a more profound understanding of individualized visual behavior.

3.3. Adaptive Fixation Prioritization

While the observer-centric feature integration focuses on the fusion of input features, the adaptive fixation prioritization module addresses the variations of gaze behavior at the output end of the decoder. To achieve this, instead of directly predicting fixation positions, our approach, aimed at individualizing fixation predictions, takes a distinct path. We start by extracting semantic feature maps, denoted as $\boldsymbol{A}_t$, from the decoder. These feature maps are subsequently prioritized using attention weights specific to each observer, providing a pragmatic means of refining fixation outputs based on their unique attention traits.

To elaborate on the process, we begin by element-wise multiplication of the semantic feature maps $\boldsymbol{A}_t$ with the input visual features $\boldsymbol{E}$, and then perform average-pooling along the spatial dimensions, resulting in a feature vector that characterizes the observer's attention distribution across different semantic feature channels, defined as

$$\boldsymbol{V}_t = \text{AvgPool}_s(\boldsymbol{E} \circ \boldsymbol{A}_t). \tag{7}$$

Considering that the visual preferences of various semantic features may vary for different observers, we introduce normalized attention weights $\boldsymbol{\beta}$ that prioritize the different feature channels, taking into account the observer feature:

$$\boldsymbol{\beta}_t = \text{softmax}(\boldsymbol{w}_b^T \tanh(\boldsymbol{W}_b\boldsymbol{V}_t + \boldsymbol{W}_{um}\boldsymbol{u})), \tag{8}$$

where $\boldsymbol{W}_b$, $\boldsymbol{W}_{um}$ and $\boldsymbol{w}_b$ are learnable parameters. Finally, the attention weights $\boldsymbol{\beta}_t$ are applied to the corresponding semantic feature maps $\boldsymbol{A}_t$ to compute the output

$$\boldsymbol{m}_t = \boldsymbol{\beta}_t^T \boldsymbol{A}_t. \tag{9}$$

This mechanism reshapes the scanpath prediction process into a weighted combination of multi-channel feature maps, allowing for the adaptive integration of these maps into the output fixation map. This approach allows the models to refine the fixation positions, providing a precise prediction of an individual's unique scanpath.

4. Experiments

This section reports comprehensive experimental results and analyses, demonstrating the effectiveness and generalizability of our method across various datasets, model architectures, and visual tasks. For further results, analyses, and implementation details, please refer to the Supplementary Material.

4.1. Experiment Settings

Tasks and Datasets. We conduct experiments on four eye-tracking datasets featuring a variety of visual tasks, including free-viewing, visual search, and visual question answering: *OSIE* [79] comprising 700 images with free-viewing gaze data from 15 undergraduate and graduate students aged 18–30, *OSIE-ASD* [71] with free-viewing gaze data from 20 individuals with ASD and 19 controls, spanning ages 21 to 60, including 33 males and 6 females, *COCO-Search18* [83] (target-present subset) featuring 6202 images with gaze data from 6 males and 4 females aged 18 to 30, collected under a visual search task, and *AiR-D* [12] offering images and questions from the GQA dataset [32] with gaze

	OSIE [79]			OSIE-ASD [71]			COCO-Search18 [83]			AiR-D [12]		
Method	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
Human	0.386	0.808	7.486	0.370	0.783	7.720	0.458	0.809	1.777	0.405	0.801	7.966
SaltiNet [2]	0.151	0.739	8.790	0.137	0.735	8.688	0.127	0.712	3.821	0.116	0.747	10.661
PathGAN [3]	0.056	0.744	9.393	0.042	0.732	9.342	0.231	0.714	2.454	0.072	0.739	9.888
IOR-ROI [69]	0.294	0.791	7.966	0.301	0.788	7.655	0.197	0.787	7.087	0.239	0.791	8.584
ChenLSTM [14]	0.373	0.804	7.309	0.341	0.791	7.602	0.454	0.799	1.932	0.356	0.808	7.845
Gazeformer [51]	0.372	0.809	7.298	0.388	0.792	7.081	0.432	0.796	2.023	0.349	0.810	8.004
ChenLSTM-FT	0.378	0.808	7.344	0.394	0.796	7.067	0.454	0.804	1.936	0.341	0.806	8.282
Gazeformer-FT	0.373	0.810	7.319	0.387	0.795	7.083	0.432	0.796	2.026	0.350	0.812	8.068
ChenLSTM-ISP	0.377	0.810	7.284	0.401	**0.798**	**6.599**	**0.480**	**0.811**	**1.862**	**0.371**	0.813	**7.651**
Gazeformer-ISP	**0.390**	**0.813**	**7.163**	**0.406**	0.797	6.823	0.455	0.806	1.997	0.362	**0.814**	7.911

Table 1. Comparison of value-based evaluation results for models' ability to predict the scanpaths of individual observers.

and question-answering data from 16 males and 4 females aged 18 to 38. Dataset splits follow ChenLSTM [14] for the OSIE, OSIE-ASD, and AiR-D datasets, and the Gazeformer [51] for the COCO-Search18.

Evaluation Metrics. We conduct individualized scanpath prediction evaluation using two complementary sets of metrics: value-based metrics and ranking-based metrics. The **value-based** metrics measure the similarity or dissimilarity between the prediction and ground-truth scanpaths of the same observer. Different from existing studies [14] that compare a generic prediction with all observers' ground-truth scanpaths, we evaluate each individualized prediction against the corresponding observer's ground truth. Specifically, *ScanMatch (SM)* [16, 65] measures the similarity of fixation position and duration using the Needleman-Wunsch algorithm [5]; *MultiMatch (MM)* [21] measures scanpath similarity regarding shape, direction, length, position, and duration; *String-Edit Distance (SED)* [7, 23, 26] converts scanpaths into strings by associating each image region with a character. To evaluate how well the model predicts distinctly different scanpaths for different observers, we also employ **ranking-based** metrics. For each predicted scanpath, we rank the ground-truth scanpaths based on their ScanMatch similarity. *Recall at K (R@K)* [10, 76] quantifies whether the correct scanpath (*i.e.*, that from the same observer) is within the top-K most similar scanpaths. *Mean Reciprocal Rank (MRR)* [17, 18, 82] measures the quality of the ranking by calculating the reciprocal of the rank of the correct scanpath. Thus, the combination of value-based metrics focusing on the specific observer and ranking-based metrics considering all observers offers a comprehensive and robust performance evaluation.

Compared Models. We implement two individualized scanpath prediction models representing typical autoregressive and non-autoregressive sequential processing paradigms, respectively: *ChenLSTM-ISP* adapts the ChenLSTM [14] model, incorporating the observer encoder and the observer-centric feature integration for input processing. The model's LSTM decoder outputs are further modified for the proposed adaptive fixation prioritization. Similarly, we implement the *Gazeformer-ISP* model upon the Gazeformer [51] architecture. It replaces the original visual-semantic joint embedding with our observer-centric feature integration and changes the Transformer decoder outputs from fixation coordinates to feature maps. We compare these ISP models with their general counterparts and other general scanpath prediction models, including SaltiNet [2], PathGAN [3], and IOR-ROI [69]. In addition, we fine-tune the general models on individual observer data (*i.e.*, ChenLSTM-FT, Gazeformer-FT) to provide a baseline for assessing the impact of explicitly incorporating observer-specific characteristics.

Implementation Details. We implement ChenLSTM [14] and Gazeformer [51] following the original methods, such as using the same visual encoder (*i.e.*, ResNet-50 [29]) and task encoder (*i.e.*, RoBERTa [48] or AiR-M [12] or CenterNet [86] object detector). For both models, the number of output feature channels for $\boldsymbol{A}_t$ is empirically set to 4. Specifically, for ChenLSTM [14] and Gazeformer [51], we adopt supervised learning for 15 epochs and self-critical sequence training (SCST) [14, 60] for the remaining 10 epochs. In supervised learning, we train our model using the Adam [42] optimizer with learning rate 10^{-4} and weight decay 5×10^{-5}, while in the SCST, we linearly decayed learning rates starting at 10^{-5}. To improve the learning of discriminative features across observers, each training batch includes different scanpaths for the same image.

4.2. Quantitative Results

We present value- and ranking-based evaluation results to assess the effectiveness of our ISP models in capturing the unique attention traits of individual observers.

Table 1 presents the **value-based** evaluation results revealing how model predictions resemble the ground truth scanpath of each observer. While fine-tuning leads to minor

	OSIE [79]			OSIE-ASD [71]			COCO-Search18 [83]			AiR-D [12]		
Method	MRR ↑	R@1 ↑	R@5 ↑	MRR ↑	R@1 ↑	R@5 ↑	MRR ↑	R@1 ↑	R@5 ↑	MRR ↑	R@1 ↑	R@5 ↑
SaltiNet [2]	0.213	5.619	32.286	0.107	2.454	12.454	0.293	10.114	49.804	0.295	10.210	49.930
PathGAN [3]	0.221	6.667	33.048	0.110	2.601	12.894	0.294	10.082	50.245	0.293	10.000	50.629
IOR-ROI [69]	0.218	6.762	31.524	0.109	2.784	12.454	0.292	9.673	50.507	0.291	9.814	48.567
ChenLSTM [14]	0.222	7.048	32.952	0.108	2.418	13.114	0.296	10.199	50.719	0.297	9.957	51.433
Gazeformer [51]	0.223	7.048	32.476	0.107	2.564	11.758	0.292	9.873	50.114	0.299	10.459	51.361
ChenLSTM-FT	0.225	6.667	34.381	0.113	2.711	12.637	0.298	10.641	49.820	0.294	10.118	50.262
Gazeformer-FT	0.217	6.000	32.857	0.108	2.528	13.223	0.293	10.183	50.000	0.300	9.599	51.863
ChenLSTM-ISP	**0.291**	**12.667**	**44.095**	**0.147**	**4.835**	**19.194**	**0.369**	**16.639**	**61.769**	**0.338**	**13.610**	57.235
Gazeformer-ISP	0.268	10.095	41.905	0.141	4.286	18.571	0.353	15.299	60.020	0.334	13.539	**57.450**

Table 2. Comparison of ranking-based evaluation results for models' ability to distinguish different observers.

Modules			ChenLSTM						Gazeformer					
OE	FI	FP	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
			0.341	0.791	7.602	0.108	2.418	13.114	0.388	0.792	7.081	0.107	2.564	11.758
✓			0.377	0.791	7.112	0.110	2.601	13.000	0.397	0.796	7.079	0.122	3.017	15.092
✓	✓		0.389	0.795	7.064	0.122	3.150	15.238	0.398	0.796	6.982	0.134	3.810	17.509
✓		✓	0.389	0.795	7.063	0.112	2.784	13.150	0.397	0.797	7.073	0.120	3.077	15.165
✓	✓	✓	**0.401**	**0.798**	**6.599**	**0.147**	**4.835**	**19.194**	**0.406**	**0.797**	**6.823**	**0.141**	**4.286**	**18.571**

Table 3. Ablation study for the proposed technical components: observer encoder (OE), observer-centric feature integration (FI), and adaptive fixation prioritization (FP).

improvements in some cases (*e.g.*, OSIE and OSIE-ASD), it struggles on datasets with less distinct inter-observer differences (*e.g.*, COCO-Search18 and AiR-D). In contrast, the ISP models consistently outperform the general methods and fine-tuning, indicating their ability to adapt to the unique attention traits of observers. This is particularly evident in the improved performance (*e.g.*, Gazeformer-ISP, SM=0.406) on the OSIE-ASD dataset with a diverse range of observer demographics. These results suggest that our method, by directly targeting the modeling of observer-specific attention patterns, offers more robust and effective individualization.

Table 2 presents **ranking-based** evaluation comparing models' ability to distinguish ground-truth scanpaths. General models, which are observer-agnostic, cannot differentiate the ground-truth scanpaths from similar ones (*e.g.*, ChenLSTM, R@1=2.4% on OSIE-ASD, lower than random). Even after fine-tuning with individual eye-tracking data, their performance improvements are marginal (*e.g.*, ChenLSTM-FT, R@1=2.7% on OSIE-ASD), because independently tuning parameters cannot effectively learn features that distinguish each observer from the others. Differently, the individualized models achieve promising results across all metrics and datasets. From ChenLSTM to ChenLSTM-ISP, R@1 is significantly improved to 4.8% on the OSIE-ASD dataset, doubling the probability of finding the correct scanpath. It suggests that the ISP models can predict scanpaths that align closely with an observer's unique attention traits. Between network architectures, ChenLSTM-ISP consistently outperforms Gazeformer-ISP when ranking scanpaths. This performance gain may be attributed to LSTM's autoregressive nature which is more effective than Transformer's parallel approach in learning fine-grained spatiotemporal differences.

4.3. Ablation Study

To evaluate the significance of the three technical components: observer encoder (OE), observer-centric feature integration (FI), and adaptive fixation prioritization (FP), we conduct an ablation study on the OSIE-ASD dataset [79] by applying them incrementally to the ChenLSTM and Gazeformer models. Table 3 shows that a fundamental module OE results in a significant improvement in the value-based evaluation and highlights its role of encoding attention traits of observers. Furthermore, based on OE, both FI and FP have notable impacts on the model performance. First, both components achieve similar performance improvements in SM, MM, and SED, demonstrating their ability to improve the overall accuracy of scanpath predictions. Further, regarding the MRR, R@1, and R@5 metrics, FI results in more significant improvements than FP, suggesting that the seamless integration of various input features is more substantial than FP's ability to prioritize where to look at the output end. We also notice that combining both modules leads to the most significant overall performance improvements, indicating that FI and FP offer complementary enhancements. Ablation studies on the other datasets are reported in the Supplementary Material.

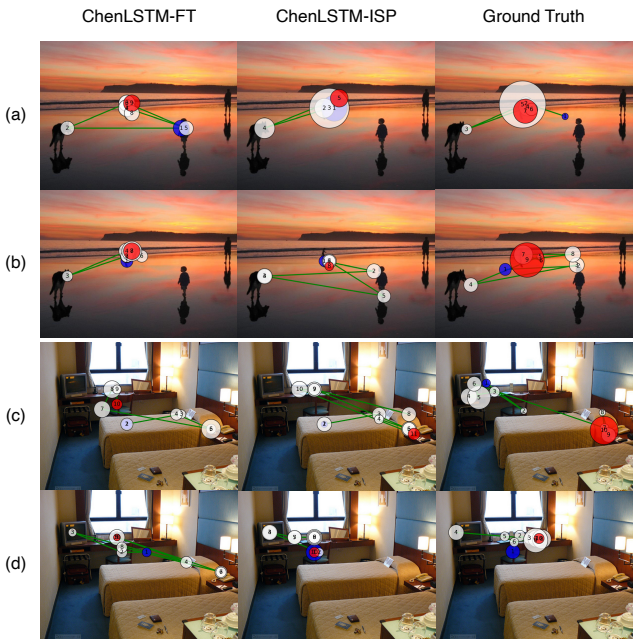

Figure 3. Qualitative examples of scanpaths predicted by ChenLSTM-FT, ChenLSTM-ISP, and ground truth. Each row compares the model predictions and the ground truth scanpath of one observer. These observers show different gaze patterns, including (a) focusing on the image center, (b) exploring different people and objects, (c) exploring broadly in the scene, and (d) focusing on a particular region. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

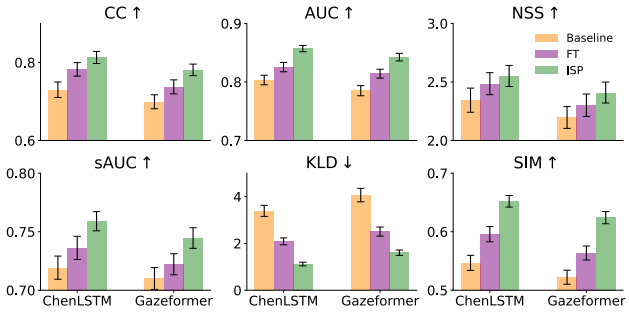

Figure 4. Saliency evaluation results of the baselines, fine-tuned (FT) models, and ISP models. Error bars indicate the standard error of the mean.

4.4. Qualitative Examples

To understand how the predicted scanpaths align with observer-specific gaze patterns, we present a qualitative comparison in Figure 3. Figure 3a and Figure 3b compare the scanpaths between an observer with autistic traits and a non-autistic observer. It can be seen that observer (a) focused on the center of the image while avoiding direct gaze at people, while observer (b) looked at people more frequently. Figure 3c and Figure 3d compare the scanpaths of two observers responding to the question 'What is the device on top of the nightstand made of wood?' with different answers. Observer (c) successfully found the correct answer 'phone' by searching broadly within the image, but observer (d) responded with an incorrect answer 'television' because the fixations were mostly distributed around the television. Notably, while the fine-tuning approach (column 1) falls short in capturing observer-specific gaze patterns, the ISP models' predictions (column 2) better align with the scanpaths of the human observers (column 3). This capability of ISP models opens up new avenues for understanding and interpreting individual differences in visual perception and decision-making processes.

4.5. From Scanpaths to Saliency Maps

To further confirm the effectiveness of our ISP method, we assess the spatial accuracy of the predicted fixations using established saliency evaluation metrics [31, 37], including Linear Correlation Coefficient (CC), Area Under the ROC curve (AUC), Normalized Scanpath Saliency (NSS), shuffled AUC (sAUC), Kullback-Leibler divergence (KLD), and similarity metric (SIM). Saliency maps are generated by aggregating predicted fixations from all observers and applying a Gaussian kernel smoothing to all fixation points. Figure 4 shows the substantial improvement of the ISP models over the baselines and fine-tuned models when applied to the OSIE-ASD [71] dataset. This improvement shows that our method not only accurately predicts individual observers' fixations but also enhances the overall prediction of fixation distributions for the population.

4.6. Semantic Analyses

Moving forward, we conduct statistical analyses on the OSIE-ASD dataset to test ISP models' ability to learn the attention differences across observers and populations. While the evaluations above focus on fixation positions and durations, this analysis considers how the predicted fixations align with the ground truth regarding their semantic-level statistics. Specifically, we group fixations into three categories based on the region of interest (ROI) annotations provided by OSIE [79], which are social regions (directly relating to humans, including faces, emotion, touched, gazed), nonsocial regions (*e.g.*, implied motion, relating to nonvisual senses, designed to attract attention, and other objects), and background. Each observer has a unique fixation distribution over the three categories (*i.e.*, social, nonsocial, and background), which enables the following individual-level and population-level analyses.

Individual Level. To evaluate how the predicted scanpaths resemble human fixation statistics, we rank observers by their proportion of fixations in each category. The fixations can be obtained from the model predictions or the ground truth. Table 4 presents Spearman's rank correlation

Method	Social	Nonsocial	Background
ChenLSTM [14]	0.181	-0.159	0.067
Gazeformer [51]	-0.141	-0.253	-0.211
ChenLSTM-FT	0.137	0.040	-0.166
Gazeformer-FT	0.045	0.164	0.051
ChenLSTM-ISP	**0.621**	**0.655**	**0.720**
Gazeformer-ISP	**0.692**	**0.572**	**0.699**

Table 4. Spearmans' correlation coefficients of fixation proportions in 3 semantic ROIs (*i.e.*, social, nonsocial, and background) between the ground truth and predictions. Bold numbers indicate significant positive correlations ($p < 0.05$).

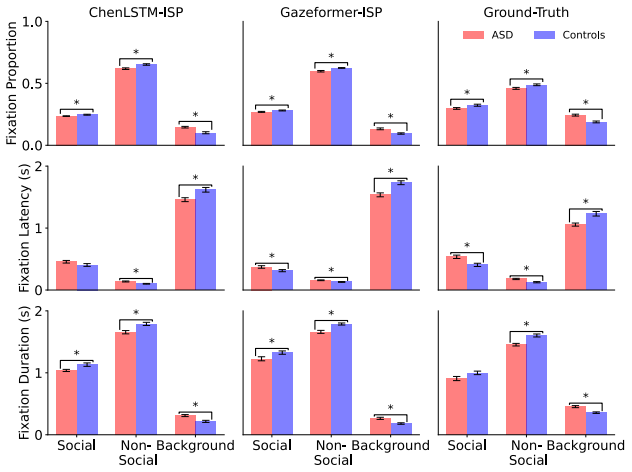

Figure 5. Statistical comparison between the predicted fixations for the ASD and Control groups [71]. Error bars indicate the standard error of the mean. Asterisks indicate significant differences (unpaired t-test, $p < 0.05$).

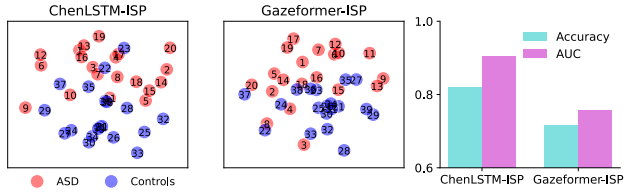

Figure 6. Visualization of features extracted from ISP models (numbers indicate observer identities) and results of ASD classification using the features.

coefficient [70] to compare the observer rankings between the predictions and the ground-truth fixations. While fine-tuning is less effective, showing low correlations across all categories, ISP models consistently achieve significant and high positive correlations, suggesting their ability to resemble each human observer's unique fixation patterns.

Population Level. Beyond individual characterization, ISP models also effectively capture and reproduce distinctive attention traits observed at the population level. For example, individuals with ASD exhibit lower proportions, higher latency, and shorter duration of fixations to both social and nonsocial cues [45, 66, 71]. Figure 5 shows that fixations predicted by the ISP models achieve similar statistics. The statistical agreement between the model predictions and the ground-truth scanpaths demonstrates our method's ability to generalize and represent population-level characteristics, reinforcing its potential utility in a variety of applications.

4.7. Application

To showcase the potential applicability of ISP models in the diagnosis of neurodevelopmental conditions, we visualize ISP model features and use these features to classify people with ASD. First, the individualization ability of our method is highlighted through t-distributed stochastic neighbor embedding (t-SNE) visualization. By concatenating all observer-specific features from Equations (2), (4), (5), and (8), into $\boldsymbol{v} = [\boldsymbol{W}_{mu}\boldsymbol{u} \,\|\, \boldsymbol{W}_{us}\boldsymbol{u} \,\|\, \boldsymbol{W}_{uc}\boldsymbol{u} \,\|\, \boldsymbol{W}_{um}\boldsymbol{u}]$, where $\|$ represents the vector concatenation, Figure 6 shows that the ISP model features can clearly distinguish people with ASD from the controls. It is noteworthy that such features are learned in an unsupervised manner without knowing each observer's class label, suggesting the strong learning power of the ISP models. Further, based on a leave-one-out cross-validation, we train a two-layer perceptron to classify people with ASD using the extracted feature $\boldsymbol{v}$. ChenLSTM-ISP and Gazeformer-IPS achieve 82.1% and 71.8% classification accuracy, respectively, similar to clinical gold standards [24, 36]. These results demonstrate ISP models' potential in real-world healthcare applications.

5. Conclusion

We have introduced a novel approach to predicting individualized human visual scanpaths. Our approach features three novel components: observer encoder, observer-centric feature integration, and adaptive fixation prioritization. Through extensive experiments across multiple datasets, network architectures, and visual tasks, our method consistently outperforms state-of-the-art scanpath prediction methods and individualization based on observer-specific fine-tuning. The results demonstrate the method's ability to generate human-like scanpaths and account for individual observers' gaze patterns. By providing a better understanding of how individuals process visual information, our study has significant implications for tailored, user-centric solutions, such as improving the design of interfaces, products, and services across a wide range of application domains.

Acknowledgments

This work is supported by NSF Grant 2143197.

References

[1] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye gaze techniques for human computer interaction: A research survey. *Virtual Reality*, 2023. 1

[2] Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto, and Noel E. O'Connor. SaltiNet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017. 5, 6

[3] Marc Assens, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018. 5, 6

[4] Bahar Aydemir, Ludo Hoffstetter, Tong Zhang, Mathieu Salzmann, and Sabine Susstrunk. TempSAL - uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[5] Saul B.Needleman and Christian D.Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology (JMB)*, 1970. 5

[6] Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581v1*, 2015. 2

[7] Stephan A. Brandt and Lawrence W. Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience (JCN)*, 1997. 5

[8] Patrick Le Callet and Ernst Niebur. Visual attention and applications in multimedia technologies. *Proceedings of the Institution of Electrical Engineers*, 2013. 1

[9] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia (TMM)*, 2022. 2

[10] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[11] Shi Chen and Qi Zhao. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[12] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 4, 5, 6

[13] Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He. Learning from unique perspectives: User-aware saliency modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[14] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 5, 6, 8

[15] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing (IEEE TIP)*, 2018. 2

[16] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods (BRM)*, 2010. 5

[17] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[18] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019. 5

[19] Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 2

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[21] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods (BRM)*, 2012. 5

[22] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *ACM Multimedia Systems Conference (MMSys)*, 2019. 1, 2

[23] Lapo Faggi, Alessandro Betti, Dario Zanca, Stefano Melacci, and Marco Gori. Wave propagation of visual stimuli in focus of attention. *arXiv preprint arXiv:2006.11035*, 2020. 5

[24] Torbjörn Falkmer, Katie Anderson, Marita Falkmer, and Chiara Horlin. Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European Child & Adolescent Psychiatry*, 2013. 8

[25] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. Predicting visual importance across graphic design types. In *ACM Symposium on User Interface Software and Technology*, 2020. 2

[26] Tom Foulsham and Geoffrey Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision (JoV)*, 2008. 5

[27] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia (TMM)*, 2016. 1

[28] Xinyue Gui, Koki Toda, Stela Hanbyeol Seo, Chia-Ming Chang, and Takeo Igarashi. "I am going this way": Gazing eyes on self-driving car show multiple driving directions. In *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2022. 1

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 3

[31] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 7

[32] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[33] Thomas E. Hutchinson, K. Preston White, Worthy N. Martin, Kelly C. Reichert, and Lisa A. Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics (TSMC)*, 1989. 1

[34] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 1998. 2

[35] Sen Jia and Neil D. B. Bruce. EML-NET:an expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 2020. 2

[36] Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[37] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 7

[38] Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. Fantastic answers and where to find them: Immersive question-directed visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[39] Ming Jiang, Sunday M Francis, Angela Tseng, Diksha Srishyla, Megan DuBois, Katie Beard, Christine Conelea, Qi Zhao, and Suma Jacob. Predicting core characteristics of asd through facial emotion recognition and eye tracking in youth. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020. 1

[40] Yue Jiang, Luis A. Leiva, Hamed R. Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. UEyes: Understanding visual saliency across user interface types. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023. 1, 2

[41] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[43] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 2

[44] Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision (JoV)*, 2022. 2

[45] Mark H. Lewis and James W. Bodfish. Repetitive behavior disorders in autism. *Developmental Disabilities Research Reviews*, 1998. 2, 8

[46] Aoqi Li and Zhenzhong Chen. Individual trait oriented scanpath prediction for visual attention analysis. In *IEEE International Conference on Image Processing (ICIP)*, 2017. 2

[47] Leida Li, Yu Zhou, Weisi Lin, Jinjian Wu, Xinfeng Zhang, and Beijing Chen. No-reference quality assessment of deblocked images. *Neurocomputing*, 2016. 1

[48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[49] Xinhui Luo, Zhi Liu, Weijie Wei, Linwei Ye, Tianhong Zhang, Lihua Xu, and Jijun Wang. Few-shot personalized saliency prediction using meta-learning. *Image and Vision Computing*, 2022. 2

[50] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research (VR)*, 2015. 2

[51] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5, 6, 8

[52] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention. *IEEE International Conference on Consumer Electronics*, 2020. 2

[53] Felix Joseph Mercer Moss, Roland Baddeley, and Nishan Canagarajah. Eye movements to natural images as a function of sex and personality. *PLoS One*, 2012. 1

[54] Young Hoon Oh and Da Young Ju. Age-related differences in fixation pattern on a companion robot. *Sensors*, 2020. 1

[55] Uchenna Chinyere Onyemauche, Samuel Makuochi Nkwo, and Charity Elochukwu Mbanusi. Towards the use of eye gaze tracking technology: Human computer interaction (hci) research. In *African Human-Computer Interaction Conference: Inclusiveness and Empowerment*, 2021. 1

[56] Matthew F. Peterson and Miguel P. Eckstein. Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, 2013. 1

[57] Thammathip Piumsomboon, Gun Lee, Robert W. Lindeman, and Mark Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *IEEE Symposium on 3D User Interfaces (3DUI)*, 2017. 1

[58] Kun Qian, Tomoki Arichi, Anthony Price, Sofia Dall'Orso, Jonathan Eden, Yohan Noh, Kawal Rhode, Etienne Burdet, Mark Neil, A. David Edwards, and Joseph V. Hajnal. An eye tracking based virtual reality system for use inside magnetic resonance imaging systems. *Scientific Reports*, 2021. 1

[59] Mengyu Qiu, Yi Guo, Mingguang Zhang, Jingwei Zhang, Tian Lan, and Zhilin Liu. Simulating human visual system based on vision transformer. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, 2023. 2

[60] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[61] Evan F. Risko, Nicola C. Anderson, Sophie Lanthier, and Alan Kingstone. Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*, 2012. 1

[62] Negar Sammaknejad, Hamidreza Pouretemad, Changiz Eslahchi, Alireza Salahirad, and Ashkan Alinejad. Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in Cognitive Psychology*, 2017. 1

[63] Bahman Abdi Sargezeh, Niloofar Tavakoli, and ohammad Reza Daliri. Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiology & Behavior*, 2019. 1

[64] Anjana Sharma and Pawanesh Abrol. Eye gaze techniques for human computer interaction: A research survey. *International Journal of Computer Applications*, 2013. 1

[65] Hiroyuki Sogo. Gazeparser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior Reserch Methods (BRM)*, 2013. 5

[66] Mikle South, Sally Ozonoff, and William M. McMahon. Repetitive behavior profiles in asperger syndrome and high-functioning autism. *Journal of Autism and Developmental Disorders*, 2005. 2, 8

[67] Tommy Strandvall. Eye tracking in human-computer interaction and usability research. In *IFIP Conference on Human-Computer Interaction*, 2009. 1

[68] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. ScanDMM: A deep markov model of scanpath prediction for 360° images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[69] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019. 2, 5, 6

[70] Alexander Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2011. 8

[71] Shuo Wang, Ming Jiang, Xavier Morin, Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 2015. 2, 4, 5, 6, 7, 8

[72] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[73] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cognitive Processing (CP)*, 2017. 2

[74] Ze-Yu Wang and Ji Young Cho. Older adults' response to color visibility in indoor residential environment using eye-tracking technology. *Sensors*, 2022. 1

[75] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[76] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[77] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian Conference on Computer Vision (ACCV)*, 2018. 1

[78] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 1

[79] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision (JoV)*, 2014. 2, 4, 5, 6, 7

[80] Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao. Beyond universal saliency: Personalized saliency prediction with multi-task cnn. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2017. 2

[81] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2018. 2

[82] Jinhui Yang, Xianyu Chen, Ming Jiang, Shi Chen, Louis Wang, and Qi Zhao. VisualHow: Multimodal problem solving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[83] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh

Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 5, 6

[84] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[85] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383*, 2023. 2

[86] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5