LEARNING SHARED VISUAL EXPLORATION PRIMITIVES FOR CROSS-SUBJECT MENTAL DISORDER DIAGNOSIS FROM EYE MOVEMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject mental disorder diagnosis from eye movements requires learning disease-discriminative representations that generalize beyond the subjects observed during training. This is challenging because eye-movement behaviors exhibit substantial inter-subject variability, such that the same disorder may manifest through heterogeneous visual exploration strategies. Existing methods may collapse individual-specific pathological manifestations into a common representation, limiting generalization to unseen subjects. Motivated by shared structural patterns in eye movements across individuals, we propose to represent eye movements using shared, fine-grained, latent visual exploration primitives and their personalized compositions. Thus, we develop a shared Visual Exploration Primitive-Aware (VEPA) model for cross-subject mental disorder diagnosis. By jointly modeling shared primitives and personalized compositions, VEPA learns robust, disease-discriminative representations that generalize to unseen subjects. Specifically, our model first employs a Visual Primitive Discovery module (VPD) to model diverse visual exploration primitives as Gaussian distributions, whose correspondence with eye-movement patterns is characterized through post-hoc visualization. Then a Personalized Primitive Composition module (PPC) is designed to efficiently encode primitive compositions via differentiable Gumbel-Softmax routing and adaptive layer selection, with contrastive learning boosting classification. Experiments on two clinical eye-tracking datasets for depression and schizophrenia diagnosis demonstrate that our model outperforms existing methods, achieving a 2.3% relative accuracy improvement.

1 INTRODUCTION

Eye tracking with image stimuli captures individual attentional preferences and enables large-scale mental disorder classification, facilitating early intervention and health management Li et al. (2024); Liu et al. (2025a). Inter-subject variability induces high heterogeneity in eye-movement preferences Wu et al. (2026); Zhu et al. (2025), making this a challenging cross-subject generalization problem: disease-discriminative representations learned from the training set (observed subjects) often fail to generalize to unseen test subjects.

Existing cross-subject methods primarily consist of adaptation-based (e.g., domain adaptation, transfer learning, meta-learning) and generalization-based categories (e.g., disentanglement, prototype learning) Ding et al. (2025). Adaptation-based methods require unlabeled samples from unseen subjects during training Jiménez-Guarneros & Fuentes-Pineda (2024); Liu et al. (2024; 2025b), or even extra fine-tuning Gong et al. (2025); Liu et al. (2026), limiting their practicality and generalization, particularly in real-time assessment. Generalization-based methods learn subject-invariant representations directly, enabling generalization to unseen subjects without additional adaptation Jia et al. (2024); Zhao et al. (2025); Cai & Pan (2023); Hu et al. (2025); Wang et al. (2024).

However, as illustrated in Figure 1, these methods typically remove inter-subject variations through alignment or disentanglement, which may force distinct pathological manifestations across individuals into a common representation and limit generalization to unseen subjects. Recent studies have characterized scanpath behavior in eye movements through shared structural patterns across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

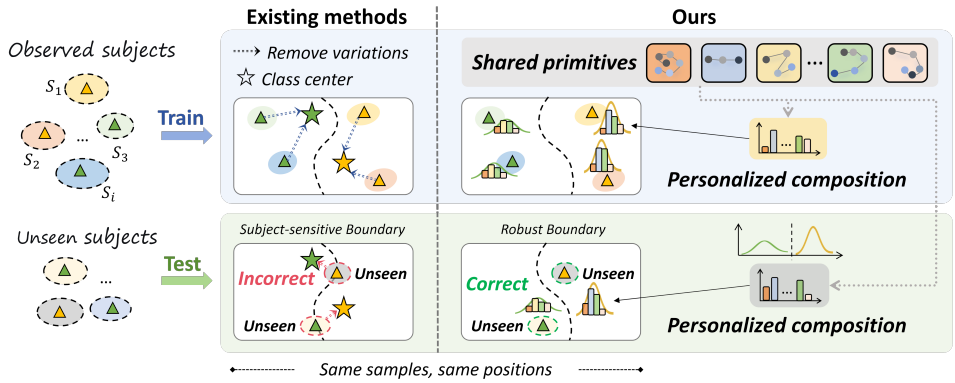


Figure 1: Existing methods vs. Our VEPA model.

individuals Laborde et al. (2026), such as refixation Nikolaev et al. (2018) and focal viewing Guo et al. (2022). Inspired by this shared-structure perspective, rather than explicitly modeling specific eye-movement behaviors, we adopt a shared-structure representation of eye movements and learn a set of fine-grained latent visual exploration primitives shared across individuals. We further hypothesize that individuals selectively compose these primitives into personalized behaviors, which may reflect heterogeneous pathological manifestations. Therefore, learning such latent primitives from observed subjects provides a shared representational basis for generalization to unseen subjects.

Based on this insight, we learn disease-discriminative representations by explicitly capturing both the shared visual exploration primitives and their personalized compositions: **a) Shared Primitives** define a shared representational basis for disease-discriminative eye-movement modeling. A set of shared, fine-grained latent visual exploration primitives captures complementary aspects of visual exploration behavior across individuals, providing more transferable representational components than holistic trajectory modeling. Existing methods directly optimize trajectory-level representations, overlooking these underlying primitives and consequently failing to learn robust representations for unseen subjects. **b) Personalized Composition** determines the representational structure. While these visual exploration primitives are shared across individuals, disease-discriminative representations emerge through their personalized selection and adaptive composition. Therefore, individual differences reflect distinct compositions of shared primitives rather than arbitrary behavioral variations. Current methods overlook this individualized compositional process, limiting their ability to adapt shared primitives to unseen subjects.

Thus, we propose a shared Visual Exploration Primitive-Aware (VEPA) model for cross-subject mental disorder diagnosis, designed to explicitly model robust disease-discriminative representations with a dual focus on shared primitives and personalized composition. Our model consists of a Visual Primitive Discovery module (VPD) and a Personalized Primitive Composition module (PPC): **a)** VPD models latent visual exploration primitives as Gaussian distributions, whose correspondence with eye-movement patterns is characterized through post-hoc visualization, while promoting primitive diversity via an orthogonality loss. Then the Wasserstein distance is employed to match eye-movement distributions with these primitives, decomposing trajectories into transferable activations of shared primitives. **b)** Based on these matches, PPC leverages a differentiable Gumbel-Softmax routing mechanism to achieve end-to-end personalized primitive selection. It facilitates composition specificity via personalization losses and efficiently encodes the selected primitive compositions after adaptive layer selection. Furthermore, contrastive learning optimizes composition representations to boost classification. Our contributions are as follows:

- We propose a shared Visual Exploration Primitive-Aware (VEPA) model for cross-subject mental disorder diagnosis, modeling robust disease-discriminative representations to improve generalization to unseen subjects.
- A Visual Primitive Discovery module (VPD) is proposed to model visual exploration primitives as Gaussian distributions, promoting diversity via orthogonality loss and matching primitives via Wasserstein distance.
- A Personalized Primitive Composition module (PPC) is designed to encode primitive compositions via a differentiable routing mechanism and adaptive layer selection, facilitating composition specificity, and is further boosted by contrastive loss.

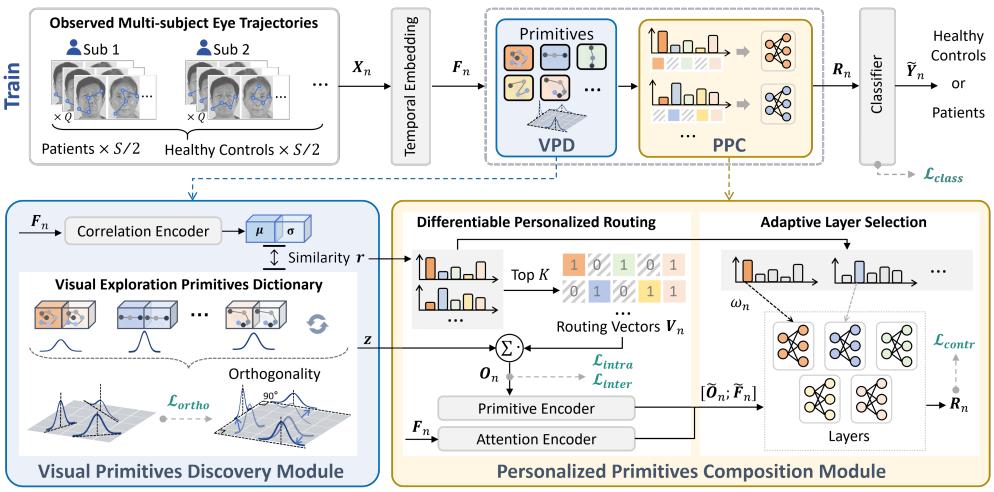


Figure 2: Schematic illustration of our VEPA model, comprising a Visual Primitive Discovery module (VPD) and a Personalized Primitive Composition module (PPC).

- Extensive experiments on two clinical eye-tracking datasets (depression and schizophrenia) demonstrate our model's effectiveness and generalizability. Visualizations of learned visual exploration primitives further provide interpretable insights into eye-movement patterns.

2 Related Work

2.1 Eye-movement Analysis Methods

Eye-movement analysis methods assist in mental disorder diagnosis by extracting potential biomarkers from eye-tracking signals, encompassing both traditional eye-movement metrics and deep learning features. Traditional metrics, such as fixation duration and saccadic amplitude, are manually defined to characterize attentional biases. For instance, depressed patients exhibit longer first fixation durations in both neutral and aggressive scenes Rantanen et al. (2021), show greater attention to negative content Akram et al. (2021), and have shorter fixation durations on positive emotional stimuli Huang et al. (2023). To capture latent characteristics of eye movements, deep learning techniques have been applied. For example, an MLP using eye-movement metrics was developed to predict depression in adolescents Lagunes-Ramírez et al. (2024). DVP Xia et al. (2022) employs a weakly supervised LSTM to capture dynamic viewing patterns for remote autism screening. EmoTransformer Li et al. (2024), based on a Transformer architecture, extracts features from stimulus images and eye-movement trajectories for depression recognition. EmMixformer Qin et al. (2025) extracts time-frequency information with a mixed Transformer, combining long-range time-dependencies and global dependencies for recognition. Fixation-guided BERT model Xia et al. (2026) incorporates an image-based pre-training model with fixation-aware encoding to extract semantic representations for classification. However, these methods overlook the cross-subject generalization problem caused by inter-subject variability, which limits their performance on unseen individuals.

2.2 Cross-subject Methods

Existing cross-subject methods are primarily categorized as adaptation-based or generalization-based Ding et al. (2025). Adaptation-based methods include domain adaptation, transfer learning, and meta-learning. Domain adaptation and transfer learning align source and target domains by adapting pre-trained models to new individuals Jiménez-Guarneros & Fuentes-Pineda (2024); Liu et al. (2024; 2025b), while meta-learning enables rapid adaptation via prior knowledge and few instances Gong et al. (2025); Liu et al. (2026). However, these methods require unlabeled samples from unseen subjects during training or additional fine-tuning on novel individual data, limiting their practicality and generalization, particularly in real-time assessment. Generalization-based approaches Jia et al. (2024); Zhao et al. (2025); Cai & Pan (2023); Hu et al. (2025); Wang et al. (2024) learn subject-invariant representations through feature disentanglement or alignment, allow-

ing the trained model to directly generalize to unseen subjects. For example, MDNet Jia et al. (2024) employs reconstruction and task constraints to decouple subject-specific and subject-invariant features. SeiFuD Zhao et al. (2025) disentangles short-time Fourier transform features via domain discrimination and adversarial losses, with reconstruction preserving information. PCDG Cai & Pan (2023) dynamically updates prototypes to construct positive and negative pairs for generalizable representation learning. CL-CS Hu et al. (2025) adopts supervised contrastive learning in a spatially-temporally-frequency encoded feature space. DMMR Wang et al. (2024) utilizes adversarial self-supervised pre-training to learn subject-invariant features, followed by supervised fine-tuning for classification. However, these methods typically remove inter-subject variations through alignment or disentanglement, which may force distinct pathological manifestations across individuals into a common representation and limit generalization to unseen subjects.

3 METHODOLOGY

Previous methods overlook that eye movements can be represented by shared, fine-grained, latent visual exploration primitives—their personalized compositions may reflect heterogeneous pathological manifestations—thereby limiting generalization to unseen subjects. Thus, we propose VEPA, a Visual Exploration Primitive-Aware model comprising a Visual Primitive Discovery module (VPD) and a Personalized Primitive Composition module (PPC) to model robust disease-discriminative representations for cross-subject mental disorder diagnosis. We first give an overview, then describe the VPD and PPC modules, and finally present the training and inference procedures.

3.1 OVERVIEW

We formulate eye-movement-based diagnosis as a binary classification task between individuals with a mental disorder and healthy controls using their eye-movement trajectories. As illustrated in Figure 2, each batch comprises $\{\boldsymbol{X}_n\}_{n=1}^{N}$, where $N = S \times Q$ corresponds to S subjects with Q samples per subject. Each sample consists of M eye-movement trajectories and is represented as $\boldsymbol{X}_n \in \mathbb{R}^{M\times L\times D}$, where L is the number of fixations and D is the input dimension, including coordinates, fixation duration, and pupil size. During training, each batch contains $S/2$ patients and $S/2$ healthy controls. For each sample $\boldsymbol{X}_n$, a BiLSTM-based Temporal Embedding module $\mathbb{E}$ first produces temporal feature $\boldsymbol{F}_n \in \mathbb{R}^{M\times L\times H}$, where H is the feature dimension. Based on $\boldsymbol{F}_n$, the Visual Primitive Discovery module ($\mathbb{G}_{VPD}$) learns a set of shared visual exploration primitives $\{\boldsymbol{z}_i \in \mathbb{R}^{H/2} \mid i = 1, \ldots, C\}$ and computes their matching similarities for each sample. The Personalized Primitive Composition module ($\mathbb{G}_{PPC}$) then generates personalized primitive composition representation $\boldsymbol{R}_n$, which is mapped by the classifier $\mathbb{FC}$ to produce final prediction $\widetilde{\boldsymbol{Y}}_n$. The overall model can be formulated as:

$$\widetilde{\boldsymbol{Y}}_n = \mathbb{FC}(\mathbb{G}_{PPC}(\mathbb{G}_{VPD}(\mathbb{E}(\boldsymbol{X}_n)), \mathbb{E}(\boldsymbol{X}_n))). \tag{1}$$

The cross-entropy loss is adopted as the training objective:

$$\mathcal{L}_{class} = -\big[\mathbf{Y}_n \log(\widetilde{\boldsymbol{Y}}_n) + (1 - \mathbf{Y}_n) \log(1 - \widetilde{\boldsymbol{Y}}_n)\big], \tag{2}$$

where $\boldsymbol{Y}_n$ denotes the ground-truth label.

3.2 VISUAL PRIMITIVE DISCOVERY MODULE

Designed to model shared, fine-grained, latent visual exploration primitives, VPD constructs a learnable primitive dictionary. Rather than using a fixed deterministic encoding, each primitive is modeled as a Gaussian distribution to accommodate variations in eye-movement representations while preserving a representative latent pattern. Specifically, for each visual primitive, we introduce learnable mean and standard deviation to capture its Gaussian representation. Suppose the dictionary contains C visual primitives with associated parameter pairs $\{(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \mid i = 1, \ldots, C\}$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ denote the mean and standard deviation of the i-th primitive, respectively. With the noise $\boldsymbol{\epsilon}_i$ sampled from a standard normal distribution, the encoding $\boldsymbol{z}_i \in \mathbb{R}^{H/2}$ of the i-th primitive is obtained via reparameterization:

$$\boldsymbol{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I). \tag{3}$$

For diverse primitive discovery, we employ an orthogonality loss to encourage diversity among visual primitives. The orthogonality loss $\mathcal{L}_{ortho}$ of C visual primitives is formulated as:

$$\mathcal{L}_{ortho} = \frac{1}{C^2} \sum_{i=1}^{C} \sum_{j=1}^{C} \left(\bar{\boldsymbol{\mu}}_i \cdot \bar{\boldsymbol{\mu}}_j - \delta_{ij}\right)^2, \tag{4}$$

where $\bar{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i / \|\boldsymbol{\mu}_i\|$ and $\bar{\boldsymbol{\mu}}_j = \boldsymbol{\mu}_j / \|\boldsymbol{\mu}_j\|$ denote the normalized mean vector of the i-th and j-th visual primitive, and $\delta_{ij} = 1_{\{i=j\}}$ is the Kronecker delta function. To match the eye-movement features with the visual primitives dictionary, we employ a Correlation Encoder to map $\boldsymbol{F}_n$ into a distributional representation. Specifically, the Correlation Encoder processes each eye movement through a sequential pipeline that consists of autocorrelation computation, timestamp-level cross-correlation across different eye-movements within the same sample, pooling, and an MLP. This process comprehensively captures intra-individual correlations, outputting the mean and standard deviation of $\boldsymbol{F}_n$. Subsequently, we introduce the Wasserstein distance to match eye-movement distributions with visual primitives, obtaining the matching similarity $\boldsymbol{r}_n = [r_{n,1}, \ldots, r_{n,C}]^\top$.

In summary, VPD learns a set of shared, fine-grained, latent visual exploration primitives, denoted as $\{\boldsymbol{z}_i \mid i = 1, \ldots, C\}$, which provides a more transferable representational basis, and yields the corresponding matching similarities for each sample.

3.3 Personalized Primitive Composition Module

To construct disease-discriminative representations through personalized selection and composition of shared primitives, PPC constructs routing vectors from matching similarity to select and combine primitives, regularizing the compositions with personalization losses. The dominant primitive determines the dedicated layer used to encode each composition. Contrastive learning further optimizes these composition representations to boost classification.

3.3.1 Differentiable Personalized Routing

To achieve personalized selection of visual primitives, we introduce a routing mechanism that selects K out of C visual primitives for each sample. Specifically, based on the matching similarities $\boldsymbol{r}_n$ between each sample $\boldsymbol{X}_n$ and the C primitives, we generate a multi-hot routing vector $\boldsymbol{V}_n \in \{0, 1\}^C$, where $V_{n,i}$ denotes the i-th entry and is set to 1 if $r_{n,i}$ is among the top-K entries of $\boldsymbol{r}_n$, and 0 otherwise. During training, we employ Gumbel-Softmax Jang et al. (2017) to enable differentiable top-K selection, while hard top-K routing is used during inference. The final primitive composition for each sample is obtained by summing the selected primitive encodings:

$$\boldsymbol{O}_n = \sum_{i=1}^{C} (V_{n,i} \boldsymbol{z}_i), \; n = 1, \ldots, N. \tag{5}$$

To ensure personalized primitive compositions, we minimize their variation within each subject and their correlations across subjects. Accordingly, we introduce two personalization losses: an intra-subject compactness loss $\mathcal{L}_{intra}$ and an inter-subject sparse correlation loss $\mathcal{L}_{inter}$. Given the S subjects in a batch, let $\mathcal{I}_s$ denote the indices of the Q samples from subject s. We compute the mean primitive composition of subject s as:

$$\boldsymbol{\Phi}_s = \frac{1}{Q} \sum_{n \in \mathcal{I}_s} \boldsymbol{O}_n, \quad s = 1, \ldots, S. \tag{6}$$

The intra-subject compactness loss penalizes variation of primitive compositions around their respective subject means:

$$\mathcal{L}_{intra} = \frac{1}{SQD_{\boldsymbol{O}}} \sum_{s=1}^{S} \sum_{n \in \mathcal{I}_s} \|\boldsymbol{O}_n - \boldsymbol{\Phi}_s\|_2^2, \tag{7}$$

where $D_{\boldsymbol{O}}$ is the dimension of $\boldsymbol{O}_n$. This loss encourages samples from the same subject to produce similar primitive compositions. To reduce cross-subject redundancy, we normalize each subject-level representation as $\bar{\boldsymbol{\Phi}}_s = \boldsymbol{\Phi}_s / \|\boldsymbol{\Phi}_s\|_2$. The inter-subject sparse correlation loss is defined as the

average absolute similarity between different subjects:

$$\mathcal{L}_{inter} = \frac{1}{S(S-1)} \sum_{\substack{s,t=1 \\ s \neq t}}^{S} \left| \bar{\boldsymbol{\Phi}}_s^\top \bar{\boldsymbol{\Phi}}_t \right|. \tag{8}$$

3.3.2 Adaptive Layer Selection

Different primitive compositions reflect distinct visual exploration behaviors and thus require dedicated encoding to preserve their compositional specificity. To this end, we design an adaptive layer selection mechanism that assigns each composition to a specialized layer based on its matching similarities $\boldsymbol{r}_n$ to the C primitives. The core idea is that the primitive with the highest similarity plays a primary role in the composition, and should thus determine which layer processes it. Specifically, we define ω_n as the index of the primary primitive for sample n, obtained by:

$$\omega_n = \arg \max_{i \in \{1,\ldots,C\}} r_{n,i}. \tag{9}$$

Based on this, we use C layers, denoted as $\{\mathbb{W}_1, \ldots, \mathbb{W}_C\}$, each implemented as an MLP. Compositions with the same ω_n are routed to the corresponding layer $\mathbb{W}_{\omega_n}$. Before being fed into the selected layer, each composition $\boldsymbol{O}_n$ is first enhanced by a Primitive Encoder—a two-layer MLP with LayerNorm and GELU after each linear layer—yielding an enhanced representation $\widetilde{\boldsymbol{O}}_n$. Meanwhile, the temporal features $\boldsymbol{F}_n$ are processed by an Attention Encoder—with 2D temporal convolution, batch normalization, GELU activation, adaptive average pooling, and multi-head spatial attention—yielding enhanced features $\widetilde{\boldsymbol{F}}_n$. These two enhanced representations are then concatenated as $[\widetilde{\boldsymbol{O}}_n; \widetilde{\boldsymbol{F}}_n]$ and fed into the selected layer $\mathbb{W}_{\omega_n}$:

$$\boldsymbol{R}_n = \mathbb{W}_{\omega_n} \left(\left[\widetilde{\boldsymbol{O}}_n; \widetilde{\boldsymbol{F}}_n \right] \right). \tag{10}$$

To further enhance discriminability, we incorporate an InfoNCE-based Oord et al. (2018) contrastive loss $\mathcal{L}_{contr}$ that pulls together intra-class features and pushes apart inter-class features. The optimized representations are classified to yield $\widetilde{\boldsymbol{Y}}_n$, with the cross-entropy loss $\mathcal{L}_{class}$ used for training. In summary, PPC produces personalized primitive composition representations, improving the model's generalization ability to adapt shared primitives to unseen subjects, thereby facilitating accurate and reliable disease classification.

3.4 Training and Inference

The model is trained in an end-to-end manner, incorporating the cross-entropy loss $\mathcal{L}_{class}$, the orthogonality loss $\mathcal{L}_{ortho}$ for primitives, the intra-subject compactness loss $\mathcal{L}_{intra}$, the inter-subject sparse correlation loss $\mathcal{L}_{inter}$, and the contrastive loss $\mathcal{L}_{contr}$. The total loss is formulated as:

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{class} + \lambda_{ortho}\mathcal{L}_{ortho} + \lambda_{intra}\mathcal{L}_{intra} \\ & + \lambda_{inter}\mathcal{L}_{inter} + \lambda_{contr}\mathcal{L}_{contr} \end{aligned} \tag{11}$$

During inference, our model processes an arbitrary number of samples in each batch for classification. VPD first matches visual primitives from the learned dictionary, after which PPC performs personalized selection and composition of these primitives, yielding robust disease-discriminative representations that generalize effectively to unseen subjects.

4 Experiments

This section presents the dataset, experimental setup, comparative results, ablation study, hyperparameter analysis, and interpretability visualizations.

4.1 Datasets

We validate our method on two clinical eye-tracking datasets.

	Methods	ACC (%)	Precision (%)	Recall (%)	PR-AUC (%)	ROC-AUC (%)	F1 (%)
ETDS	SVM	60.56±16.31	67.44±8.09	81.19±25.06	78.27±10.40	47.46±23.80	72.67±15.53
	LR	68.31±9.94	70.25±6.10	95.00±11.25	85.58±10.73	64.48±21.35	80.62±7.51
	RF	72.78±12.08	74.50±8.34	94.05±10.24	83.24±8.60	59.52±20.03	82.92±8.14
	EmoTransformer	76.62±8.13	78.71±8.21	93.26±8.95	85.00±9.85	72.88±14.79	84.87±5.44
	DVP	77.02±8.54	80.35±9.89	91.80±9.47	85.30±9.93	73.09±14.12	84.94±5.62
	EmMixformer	80.61±12.39	82.14±13.75	97.14±9.04	92.60±6.81	81.59±13.00	87.97±7.48
	FixBERT	82.16±11.83	83.02±12.05	96.56±6.85	92.38±8.22	82.80±15.00	88.73±7.29
	MDNet	78.58±10.66	81.56±13.64	95.00±8.05	87.51±14.08	71.11±28.97	86.72±6.02
	SeiFuD	78.97±11.36	90.62±10.66	79.76±10.54	90.34±12.10	79.33±21.04	84.32±8.17
	PCDG	77.33±7.05	76.20±5.83	91.90±11.62	90.96±7.06	72.90±24.38	86.38±6.69
	CL-CS	86.31±9.82	86.31±10.10	96.90±8.55	92.50±9.78	84.37±18.89	90.96±6.63
	DMMR	72.14±8.66	72.35±11.02	**98.57**±4.52	86.24±8.62	68.97±19.80	83.33±7.76
	VEPA	**88.31**±7.67	**91.35**±9.93	93.81±8.03	**97.01**±4.12	**94.09**±7.32	**91.73**±5.46
EMS	SVM	60.62±7.74	56.41±8.79	97.73±4.55	82.79±7.37	81.94±6.49	71.08±6.67
	LR	66.87±6.25	63.63±9.19	81.99±13.37	76.67±12.83	76.98±8.68	70.73±6.96
	RF	73.13±4.27	71.07±8.48	75.63±11.40	86.17±5.55	84.36±4.86	73.07±8.62
	EmoTransformer	78.75±7.77	73.60±6.99	88.07±11.64	86.89±10.85	87.04±7.68	80.11±8.65
	DVP	78.12±9.66	80.23±11.84	73.35±11.86	89.37±7.41	88.91±6.31	76.56±11.55
	EmMixformer	80.63±4.27	79.91±9.96	84.03±7.92	85.04±4.34	84.04±8.49	81.25±2.71
	FixBERT	81.25±9.68	81.64±12.02	84.09±19.81	85.90±9.78	86.77±8.14	81.13±10.43
	MDNet	77.50±3.95	80.08±11.27	76.31±14.18	82.69±6.70	78.73±6.06	76.67±5.77
	SeiFuD	78.75±3.23	78.23±7.23	80.68±7.54	85.79±4.09	82.74±4.19	79.02±2.73
	PCDG	76.25±5.20	81.28±5.07	68.04±9.87	84.57±4.58	81.96±3.54	73.76±6.59
	CL-CS	81.87±3.75	84.84±9.64	79.46±8.87	89.84±3.77	86.53±3.49	81.39±2.79
	DMMR	73.75±3.50	81.85±7.54	62.30±14.24	83.87±7.57	81.23±5.90	69.46±7.65
	MSNet*	83.13	85.75	80.51	-	89.72	82.44
	VEPA	**84.38**±3.27	**86.84**±8.44	**85.23**±7.94	**90.14**±5.76	**89.74**±4.01	**83.93**±3.51

Table 1: Comparisons with SOTA methods on ETDS and EMS. All metrics are reported as mean±std across all folds except MSNet*, for which results are reported from the original paper.

ETDS is an eye-tracking dataset for depression diagnosis, collected from a hospital under the supervision of psychiatrists. This dataset includes 65 patients with major depressive disorder (MDD) meeting DSM-5 criteria (baseline HAMD-17 score $\geq$ 17) and 27 healthy controls with no psychiatric history. All participants provided written informed consent. Each participant completed a free-viewing task with 60 facial image stimuli Yang et al. (2020), and eye movements were recorded using an Eyelink 1000 Plus eye tracker. Raw eye-movement data were preprocessed after subject-level splitting to remove outliers and interpolate missing values, producing sequences of $L = 8$ fixations with $D = 4$ features: x- and y-coordinate, duration, and pupil size. No samples from the same subject were shared across training and testing. This yielded 5,520 eye-movement trajectories (92 participants × 60 trials).

EMS Song et al. (2024) is a public eye-tracking dataset for schizophrenia diagnosis, consisting of 160 labeled subjects (80 patients and 80 healthy controls). Each participant's eye-movements were recorded during free viewing of 100 image stimuli, totaling 16,000 trajectories. Following the preprocessing pipeline in the original work Song et al. (2024), each trajectory has $L = 14$ fixations, with $D = 4$ features: x- and y-coordinate, fixation duration, and pupil size.

4.2 Experimental Setup

Experiments followed a subject-independent evaluation protocol to assess generalization to unseen subjects. For ETDS, each batch contained $S = 8$ subjects and $Q = 16$ samples per subject, extracted using overlapping sliding windows of $M = 30$ eye-movement trajectories. The predicted probabilities of all samples from each subject were averaged to obtain the subject-level prediction, and results were reported over 10-fold cross-validation. For EMS, we adopted the original protocol Song et al. (2024) with $(S, Q, M) = (64, 1, 100)$ and 4-fold cross-validation. With $Q = 1$, $\mathcal{L}_{intra}$ was identically zero and thus inactive. Performance was evaluated using Accuracy, Precision, Recall, PR-AUC, ROC-AUC, and F1-score. All experiments were conducted on a single NVIDIA GeForce RTX 2080 Ti GPU. The dictionary contained $C = 8$ primitives, of which $K = 4$ were selected per sample. The loss weights were determined through experiments, yielding $(\lambda_{ortho}, \lambda_{intra}, \lambda_{inter}, \lambda_{contr}) = (1, 1, 0.01, 1)$. We employed the Adam optimizer (initial lr=3e-4), using a cosine annealing scheduler with warm restarts. For each fold, the training subjects were randomly split 8:2 into training and validation sets, with no subject overlap. The model was trained for up to 400 epochs with early stopping (patience=200), and best model on validation set was saved.

Methods	ACC (%)	Precision (%)	Recall (%)	PR-AUC (%)	ROC-AUC (%)	F1 (%)
BiLSTM	74.97±9.85	78.96±10.08	90.71±8.05	87.71±6.97	73.97±12.81	83.90±5.71
VEPA	**88.31±7.67**	**91.35±9.93**	**93.81±8.03**	**97.01±4.12**	**94.09±7.32**	**91.73±5.46**
GRU	79.33±7.55	80.43±8.84	**95.48±7.31**	94.84±3.89	85.87±9.85	86.83±4.54
VEPA-GRU	**84.72±6.68**	**86.77±8.31**	95.24±7.69	**96.31±3.32**	**90.12±8.12**	**89.96±4.35**
Mamba	75.00±11.54	85.07±14.10	82.38±14.73	89.66±8.84	69.92±26.65	82.12±9.16
VEPA-Mamba	**85.19±8.36**	**93.02±9.22**	**86.43±10.71**	**94.83±3.96**	**84.88±10.57**	**88.98±6.44**
Transformer	76.39±10.26	82.38±12.46	**88.10±9.26**	86.39±9.45	59.48±27.40	84.22±**6.28**
VEPA-Transformer	**79.25±9.24**	**87.62±11.64**	84.76±12.41	**94.38±4.17**	**83.53±11.80**	**85.07**±6.69
Stationary	71.04±5.00	76.73±**5.69**	85.55±7.28	82.30±6.05	64.34±10.55	80.59±3.65
VEPA-Stationary	**77.11±3.10**	**81.37**±8.36	**90.48±6.39**	**93.85±3.73**	**81.71±10.08**	**84.69±2.38**
Crossformer	78.22±6.12	82.03±6.85	89.17±**6.47**	88.51±6.65	77.89±**9.08**	85.18±4.47
VEPA-Crossformer	**80.61±3.78**	**84.88±5.89**	**90.48**±11.39	**95.77±4.13**	**87.82**±13.02	**86.62**±3.50
TimesNet	76.25±9.31	79.45±12.64	**93.81±8.03**	89.80±9.67	69.92±31.00	85.09±**5.31**
VEPA-TimesNet	**86.81±7.31**	**92.32±8.87**	90.71±13.11	**94.02±4.53**	**81.75±14.79**	**90.46**±5.83
Dlinear	78.36±9.82	84.36±14.73	90.48±11.39	88.14±10.50	66.71±26.92	85.73±**5.78**
VEPA-Dlinear	**86.08±9.74**	**86.17±10.65**	**97.14±6.02**	**92.51±8.87**	**81.27±22.36**	**90.92**±6.34

Table 2: ETDS results with varying Temporal Embeddings validate our VPD and PPC.

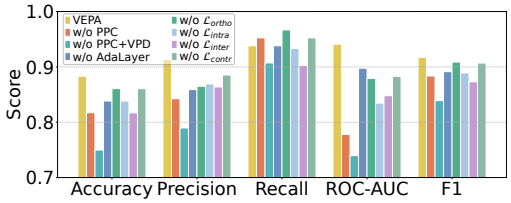
4.3 Comparison to State-of-the-art

We evaluated our method against a comprehensive set of baselines. First, we compared conventional classifiers—SVM, Logistic Regression (LR), and Random Forest (RF)—using handcrafted eye-movement metrics (fixation count, initial/total duration, saccadic amplitude, and blink percentage). We also included state-of-the-art deep learning models: a) eye-tracking-specific models, including EmoTransformer Li et al. (2024), DVP Xia et al. (2022), EmMixformer Qin et al. (2025), and FixBERT Xia et al. (2026); b) generalization-based methods designed for physiological signals, including MDNet Jia et al. (2024) SeiFuD Zhao et al. (2025), PCDG Cai & Pan (2023), CL-CS Hu et al. (2025), and DMMR Wang et al. (2024). As shown in Table 1, our VEPA achieves competitive performance across both datasets with relatively smaller standard deviations, demonstrating robust generalization to unseen subjects in each fold. On the ETDS dataset, it achieves a 2.3% relative improvement (computed as $|y - x|/x \times 100\%$) in Accuracy over CL-CS, the second-best method in terms of Accuracy. On the EMS dataset, it surpasses MSNet*, for which we report the results provided in the original paper Song et al. (2024), by 1.5% in Accuracy and 1.3% in Precision. While DMMR achieves high recall (98.57%) on ETDS, its low accuracy suggests a prediction bias rather than true discriminative power.

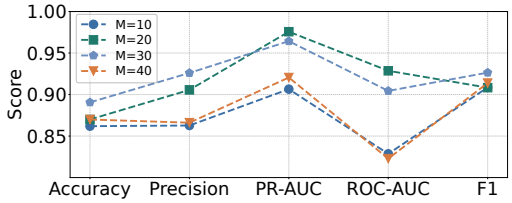
Moreover, we replace the original BiLSTM in VEPA with several time series models Wang et al. (2026) serving as the Temporal Embedding backbone: RNN-based (GRU Cho et al. (2014) and Mamba Gu & Dao (2024)), Transformer-based (Transformer Vaswani et al. (2017), Stationary Liu et al. (2022) and Crossformer Zhang & Yan (2023)), CNN-based TimesNet Wu et al. (2023), and MLP-based DLinear Zeng et al. (2023). We also evaluate these time series models individually, with consistent performance gains validating VPD and PPC. Table 2 shows that VEPA achieves the highest accuracy and one of the smallest cross-fold standard deviations, confirming its robustness.

4.4 Ablation Study and Hyperparameter Analysis

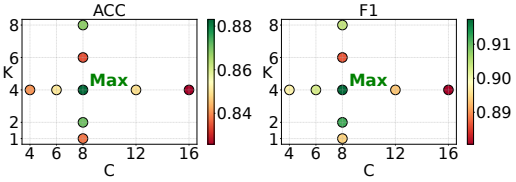
We conduct ablation experiments on the key components of VEPA, with results shown in Figure 3a, and find that removing each component leads to varying degrees of performance degradation. Furthermore, we investigate how the number of trajectories per sample affects model performance. As illustrated in Figure 3b, $M = 30$ achieves relatively strong overall performance, while accuracy remains above 0.85 across all settings, indicating limited sensitivity to trajectory count. We further evaluate sensitivity to the dictionary size $C \in \{4, 6, 8, 12, 16\}$ and selected primitive count $K \in \{1, 2, 4, 6, 8\}$. As shown in Figure 3c, $(C, K) = (8, 4)$ achieves a favorable trade-off, with Accuracy of 88.31%, and F1 of 91.73%. All configurations exceed 82% accuracy, indicating limited sensitivity to C and K. We also evaluate settings of λ_{ortho}, λ_{intra}, λ_{inter}, and λ_{contr} by varying each weight over $0.01, 0.1, 1$ while fixing the others at 1. As shown in Figure 3d, $(1, 1, 0.01, 1)$ achieves the highest PR-AUC (97.01%) and joint-highest ACC (88.31%), with competitive Precision and F1, and is adopted in our main experiments. The modest variations indicate limited sensitivity to loss weights. Moreover, two adaptive weighting strategies, Dynamic Weight Average



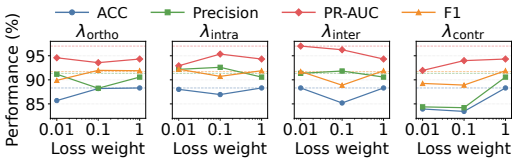
(a) Ablation on ETDS validates VEPA's components.



(b) Sensitivity of VEPA to trajectory count on ETDS.



(c) Performance comparison over C and K on ETDS.



(d) Sensitivity of VEPA to loss weights on ETDS.

Figure 3: Experimental results on ETDS.

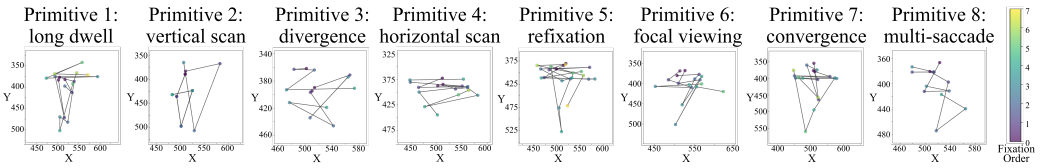


Figure 4: Behavioral interpretation of visual exploration primitives from top-activating trajectories.

(DWA) Liu et al. (2019) and GradNorm Chen et al. (2018), yield ACC/F1 scores of 86.17%/90.68% and 84.83%/88.87%, respectively, also showing VEPA's competitiveness under adaptive weighting.

4.5 Visualization Analysis

As VEPA learns latent visual exploration primitives instead of explicitly modeling specific eye-movement behaviors, we use post-hoc analysis to interpret how these primitives relate to observable patterns. We first align primitives across folds using Hungarian matching. For each primitive, we retrieve the top-3 samples with the highest matching similarities and visualize their overlaid centroid trajectories in Figure 4. Based on spatiotemporal characteristics, clinical experts provide semantic interpretations of learned latent primitives, including long dwell, vertical scan, divergence, horizontal scan, refixation, focal viewing, convergence, and multi-saccade Takemoto et al. (2023); Mańkowska et al. (2020); Li et al. (2023). These interpretations are intended to characterize the learned latent primitives rather than define explicit behavioral categories. Moreover, test-set statistics indicate that vertical scan occurs more frequently in MDD than HC (63% vs. 37%), whereas multi-saccade shows the opposite trend (40% vs. 60%), suggesting that certain learned latent primitives are associated with depression classification. Future work will further investigate the potential links between these latent primitives and disease-related processes.

5 Conclusion

In this paper, we propose VEPA, a shared visual exploration primitive-aware model for cross-subject mental disorder diagnosis. VEPA learns robust, disease-discriminative representations by jointly modeling shared primitives and their personalized compositions. Extensive experiments on two clinical eye-tracking datasets for depression and schizophrenia demonstrate the effectiveness of our method, achieving relative improvements in accuracy of 2.3% and 1.3% on ETDS and EMS. Consistent performance gains across temporal encoding backbones highlight VEPA's broad applicability, and ablation studies validate the contributions of its core components. Furthermore, hyperparameter sensitivity analyses confirm its robustness, while visualizations of learned primitives offer interpretable insights into latent visual exploration patterns relevant to mental disorder classification.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodologies or experiments, implementing methods, and assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results. Additionally, we used generative AI tools for sourcing/searching for information, editing a research paper to improve readability, and proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. Specifically, AI-generated suggestions for research methodologies and experiments were critically evaluated by the authors and only adopted when consistent with the research objectives and experimental requirements. AI-assisted code implementation was manually inspected, modified when necessary, and verified through experimental testing by the authors. AI-assisted translations and writing refinements were carefully reviewed to ensure correctness, clarity, and consistency with the original content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involves human eye-tracking data from two datasets. For the ETDS dataset in this study, all participants provided written informed consent before data collection, and the data collection procedure was conducted following the approved research protocols. All collected data were anonymized before analysis to protect participant privacy. For the publicly available EMS dataset, we used the data following the dataset's terms of use and original study protocols. No additional data collection or interaction with participants was conducted in this study. The proposed method is developed for research purposes in cross-subject mental disorder classification and is not intended to replace professional clinical diagnosis. Further validation on larger and more diverse populations is required before potential clinical applications.

REPRODUCIBILITY STATEMENT

We provide the source code and configuration files as supplementary materials for reproducing the reported results. Detailed descriptions of the VEPA architecture, training objectives, optimization procedure, experimental settings, and dataset preprocessing pipeline are provided in Sections 3–4.

REFERENCES

Umair Akram, Jason G Ellis, Glhenda Cau, Frayer Hershaw, Ashlieen Rajenthran, Mollie Lowe, Carissa Trommelen, and Jennifer Drabble. Eye tracking and attentional bias for depressive internet memes in depression. *Experimental Brain Research*, 239(2):575–581, 2021.

Honghua Cai and Jiahui Pan. Two-phase prototypical contrastive domain generalization for cross-subject eeg-based emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803, 2018.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Cheng Ding, Tianliang Yao, Chenwei Wu, and Jianyuan Ni. Advances in deep learning for personalized ecg diagnostics: A systematic review addressing inter-patient variability and generalization constraints. *Biosensors and Bioelectronics*, 271:117073, 2025.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 14247–14255, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Yuxuan Guo, Jens R Helmert, Sven-Thomas Graupner, and Sebastian Pannasch. Eye movement patterns in complex tasks: Characteristics of ambient and focal processing. *Plos one*, 17(11): e0277099, 2022.

Mengting Hu, Dan Xu, Kangjian He, Kunyuan Zhao, and Hao Zhang. Cross-subject emotion recognition with contrastive learning based on eeg signal correlations. *Biomedical Signal Processing and Control*, 104:107511, 2025.

Genying Huang, Yafang Li, Huizhong Zhu, Hong Feng, Xunbing Shen, and Zhencai Chen. Emotional stimulation processing characteristics in depression: Meta-analysis of eye tracking findings. *Frontiers in psychology*, 13:1089654, 2023.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Ziyu Jia, Fengming Zhao, Yuzhe Guo, Hairong Chen, and Tianzi Jiang. Multi-level disentangling network for cross-subject emotion recognition based on multimodal physiological signals. *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pp. 3069–3077, 2024.

Magdiel Jiménez-Guarneros and Gibran Fuentes-Pineda. Cfda-csf: A multi-modal domain adaptation method for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):1502–1513, 2024.

Quentin Laborde, Axel Roques, Allan Armougum, Nicolas Vayatis, Ioannis Bargiotas, and Laurent Oudre. Vision toolkit part 3. scanpaths and derived representations for gaze behavior characterization: a review. *Frontiers in Physiology*, 16:1721768, 2026.

Derick Axel Lagunes-Ramírez, Gabriel González-Serna, Leonor Rivera-Rivera, Nimrod González-Franco, María Y Hernández-Pérez, and José A Reyes-Ortiz. Through the youth eyes: Training depression detection algorithms with eye tracking data. *IEEE Latin America Transactions*, 23(1): 6–16, 2024.

Xin Li, Haizhuang Liu, Rongquan Wang, Bochao Zou, Yuxin Lin, and Huimin Ma. Emo transformer: Transformer-based depression detection via eye movements. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024.

Yang Li, Xiang Zhang, Xianmin Zhang, Benliang Zhu, and Xin Ye. A local optical flow eye-tracking method for depression detection. *AIP Advances*, 13(9), 2023.

Ke Liu, Xin Luo, Wenrui Zhu, Zhu Liang Yu, Hong Yu, Bin Xiao, and Wei Wu. Enhancing eeg-based cross-subject emotion recognition via adaptive source joint domain adaptation. *IEEE Transactions on Affective Computing*, 16(3):1419–1430, 2024.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1871–1880, 2019.

Xiaobo Liu, Yuxi Li, Yuan Chen, Chen Xue, Jin Fan, Jiaming Zhang, Dongling Zhong, Qinjian Dong, Zhong Zheng, Juan Li, et al. Attentional bias toward threatening stimuli in major depressive disorder: A free-viewing eye-tracking study. *Journal of Affective Disorders*, 378:352–359, 2025a.

Xuan-Hao Liu, Yan-Kai Liu, Tianyi Zhou, Bao-Liang Lu, and Wei-Long Zheng. Mindcross: Fast new subject adaptation with limited data for cross-subject video reconstruction from brain signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17589–17597, 2026.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35: 9881–9893, 2022.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fmri. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5730–5738, 2025b.

Aleksandra Mańkowska, Michał Harciarek, and Kenneth M Heilman. Feeling and looking down: impact of depressive symptoms on the allocation of vertical attention. *Cognitive and Behavioral Neurology*, 33(2):137–144, 2020.

Andrey R Nikolaev, Radha Nila Meghanathan, and Cees van Leeuwen. Refixation control in free viewing: A specialized mechanism divulged by eye-movement-related brain activity. *Journal of neurophysiology*, 120(5):2311–2324, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Huafeng Qin, Hongyu Zhu, Xin Jin, Qun Song, Mounim A El-Yacoubi, and Xinbo Gao. Emmixformer: Mix transformer for eye movement recognition. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Matti Rantanen, Jarkko Hautala, Otto Loberg, Jaakko Nuorva, Jari K Hietanen, Lauri Nummenmaa, and Piia Astikainen. Attentional bias towards interpersonal aggression in depression–an eye movement study. *Scandinavian journal of psychology*, 62(5):639–647, 2021.

Yingjie Song, Zhi Liu, Gongyang Li, Jiawei Xie, Qiang Wu, Dan Zeng, Lihua Xu, Tianhong Zhang, and Jijun Wang. Ems: A large-scale eye movement dataset, benchmark, and new model for schizophrenia recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Ayumi Takemoto, Inese Aispuriete, Laima Niedra, and Lana Franceska Dreimane. Depression detection using virtual avatar communication and eye tracking. *Journal of Eye Movement Research*, 16(2):12, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yiming Wang, Bin Zhang, and Yujiao Tang. Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 628–636, 2024.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Qianying Wu, Na Yeon Kim, and Ralph Adolphs. Using computational eye-tracking to understand individual differences in psychiatric disorders. *Nature Mental Health*, pp. 1–12, 2026.

Chen Xia, Dingwen Zhang, Kuan Li, Hongxia Li, Jianxin Chen, Weidong Min, and Junwei Han. Dynamic viewing pattern analysis: Towards large-scale screening of children with asd in remote areas. *IEEE Transactions on Biomedical Engineering*, 70(5):1622–1633, 2022.

Chen Xia, Pan Liu, Wenqi Zhong, Dingwen Zhang, and Junwei Han. Eye movement classification via global semantic representation with fixation-guided bert. *IEEE Transactions on Multimedia*, 2026.

Tao Yang, Zeyun Yang, Guangzheng Xu, Duoling Gao, Zi tong Zhang, Hui Wang, Shiyu Liu, Linfeng Han, Zhixin Zhu, Yang Tian, Yuqi Huang, Lei Zhao, Kui Zhong, Bolin Shi, Juan Li, Shimin Fu, Peipeng Liang, Michael J. Banissy, and Pei Sun. Tsinghua facial expression database – a database of facial expressions in chinese young and older women and men: Development and validation. *PLoS ONE*, 15, 2020.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Sha Zhao, Caibo Zhang, Jiquan Wang, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Seifud: A novel deep learning framework leveraging domain generalization for cross-subject seizure prediction. *Neurocomputing*, pp. 132227, 2025.

Qi Zhu, Ting Zhu, Lunke Fei, Chuhang Zheng, Wei Shao, David Zhang, and Daoqiang Zhang. Multi-modal cross-subject emotion feature alignment and recognition with eeg and eye movements. *IEEE Transactions on Affective Computing*, 2025.