Supplementary Materials: Gaze4ASD: A Novel Dataset and Visual Saliency Map-Based Method for Autism Screening

Anonymous ICME submission

I. Web Crawler Tagging for Image Collection

To obtain images that would elicit distinct eye gaze behavior between Autism spectrum disorder (ASD) and Typically developing (TD) children, we utilized a web crawler to gather a large and diverse set of images from Baidu Images (https://image.baidu.com). The web crawler used 120 carefully selected Chinese-language tags to search for images across various categories, including animals (16 tags), children's activities (30 tags), children's scenes (20 tags), outdoor and nature scenes (15 tags), transportation (9 tags), toys and playground equipment (11 tags), buildings and settings (11 tags) and others (8 tags). In total, 34,504 images were collected, ensuring a broad range of stimuli. As the images were intended for Chinese child participants, we specifically selected images featuring individuals with Asian appearances to minimize potential racial biases in the experiment.

Given that the data collection was conducted using a 2K display, the need for high-quality images was critical, as lower-resolution images would have appeared blurry during presentation. To ensure image clarity, we appended the term "high definition" to each keyword, which was essential for accurate eye gaze analysis.

The selected tags, originally in Chinese, were translated into English for clarity and are as Fig. 1.

II. Fine-tuning and Training Loss Function

As mentioned in Sec.III.B of the main paper, during Phase 2, we applied transfer learning to fine-tune the base model using data from both ASD and TD children from the Saliency4ASD dataset [1]. We adopted the same loss function used in TranSalNet [2] during the fine-tuning process. This loss function is based on commonly used visual saliency map evaluation metrics. Specifically, we employed Normalized Scanpath Saliency (NSS), Kullback-Leibler Divergence (KLD), Linear Correlation Coefficient (CC), and Similarity (SIM), which are designed to assess how well the model's predicted saliency maps match the ground-truth saliency and attention maps. The loss terms are weighted by four hyperparameters that control the relative importance of each metric during training.

The loss function is expressed as follows:

$$\begin{aligned} L(\mathbf{y}^s, \mathbf{y}^f, \hat{\mathbf{y}}) = \lambda_1 L_{\text{NSS}}(\mathbf{y}^f, \hat{\mathbf{y}}) + \lambda_2 L_{\text{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) \\ + \lambda_3 L_{\text{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) + \lambda_4 L_{\text{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}), \end{aligned} \tag{1}$$

where L represents the total loss, $\hat{\mathbf{y}}$ is the predicted saliency map generated by the model, $\mathbf{y}^s$ is the ground-truth saliency map, and $\mathbf{y}^f$ refers to the attention map. The hyperparameters λ_1, λ_2, λ_3, and λ_4 control the relative importance of the individual loss terms, allowing the model to optimize its performance by adjusting the balance between different saliency metrics during training.

III. Saliency Map Comparison for ASD and TD Children

To validate the effectiveness of our Saliency Prediction Model for ASD and TD children, we compared the predicted saliency maps with the ground truth maps from the Saliency4ASD dataset. This comparison demonstrates the model's ability to capture distinct visual attention patterns in ASD and TD children when viewing the same image stimuli. Representative examples of these comparisons are shown in Fig. 2.

The results in Fig. 2 indicate that the predicted saliency maps closely align with the ground truth, effectively capturing the differences in visual attention between ASD and TD children.

IV. Image Selection for Gaze Pattern Comparison

The following two sets of images are provided: (1) 30 images selected by our Image Selection Module, which exhibit the most significant eye gaze differences between ASD and TD children, as shown in Fig. 5a and (2) 30 control images with minimal eye gaze differences, as shown in Fig. 5b. These images provide a visual overview of the stimuli used for our eye gaze data collection.

Both sets of images were sourced from a curated dataset of over 34,504 images, which includes a wide range of categories. This dataset was further refined to exclude content unsuitable for children. The first set comprises the top 30 images ranked by the Image Selection Module based on their ability to highlight gaze differences between ASD and TD children, while the second set consists of the bottom 30 images with the smallest differences in visual saliency maps.

The top-ranked images are predominantly composed of smiling human faces in social contexts, whereas the control images largely consist of animals and non-human scenes. This distribution reflects known psychological patterns in ASD-related gaze behaviors, highlighting the module's ability to identify meaningful stimuli for analyzing gaze differences.

```
tags = [
    'cute kitten', 'cute puppy', 'elephant', 'lion', 'monkey',
    'penguin', 'giraffe', 'bird', 'butterfly', 'dolphin',
    'whale', 'turtle', 'zebra', 'tiger', 'panda',
    'rabbit', 'fairy tale castle', 'classroom', 'amusement park',
    'toy store', 'library', 'swimming pool', 'children playground',
    'children park', 'supermarket', 'train', 'bus', 'car',
    'airplane', 'boat', 'bicycle', 'toy train', 'toy car',
    'building blocks', 'toy airplane', 'rocking horse', 'swing',
    'slide', 'building blocks house', 'carousel', 'roller coaster',
    'ferris wheel', 'park gazebo', 'park bench', 'farm tractor',
    'school playground', 'church', 'bridge', 'market', 'playground swing',
    'fountain', 'tree house', 'beach', 'forest', 'ocean',
    'grass', 'blue sky and clouds', 'garden', 'waterfall', 'mountain view',
    'children playing', 'family gathering', 'school activities', 'birthday party',
    'park picnic', 'sports day', 'children hugging', 'amusement park fun',
    'kindergarten classroom', 'campus activities', 'childhood friends',
    'childhood memories', 'group games', 'outdoor activities', 'children toy house',
    'park playground equipment', 'birthday party decorations', 'family dinner',
    'sports field activities', 'library reading', 'swimming pool side',
    'beach activities', 'amusement park', 'supermarket shopping',
    'street performance', 'marketplace', 'museum visit', 'science museum',
    'smiling children', 'reading child', 'drawing child', 'running child',
    'jump rope child', 'biking child', 'sliding child', 'swinging child',
    'swimming child', 'picnic child', 'dancing child', 'building blocks child',
    'listening to story child', 'child playing with blocks', 'child reading',
    'child playing with toys', 'child drawing', 'child doing crafts',
    'child playing with sand', 'child swinging', 'child sliding',
    'child riding bicycle', 'child kicking football', 'child playing ball',
    'child blowing bubbles', 'child feeding animals', 'child using computer',
    'child watching TV', 'child playing games', 'professional elites',
    'Asian youth', 'people', 'baby', 'healthcare workers', 'education',
    'Eastern people', 'kindergarten'
]
```

Fig. 1: List of selected tags used for image collection, originally in Chinese and translated to English for clarity. Each tag was appended with the term "high definition" to ensure the retrieval of high-quality images.

V. Psychological Paradigms and Feature Extraction

To investigate differences in eye gaze behavior between ASD and TD children, we designed three psychological paradigms inspired by recent findings in eye gaze studies for ASD [3]–[5]. Data for these paradigms were collected from a subset of the recruited participants. These paradigms are designed to elicit distinct visual attention patterns through carefully structured stimuli and procedures, enabling detailed feature extraction for analysis. Below, we describe each paradigm.

A. Habituation Paradigm

This paradigm examines visual attention to repetitive and novel stimuli, assessing participants' preference for familiar versus new images. It comprises 70 trials grouped into five stimulus categories: simple shapes (non-social), clocks (complex non-social), and social stimuli, including happy, neutral, and sad facial expressions, as shown in Fig. 3. Each category contains 14 trials.

At the start of each trial, a central image is displayed to attract the child's attention. Once the child fixates on this image for 100 milliseconds, two images appear side by side: one identical to the previous trial's image (repeated stimulus) and one novel image. Each trial lasts 3 seconds, resulting in a total experiment duration of 210 seconds.

The extracted features from this paradigm include the total fixation count on each stimulus category (shapes, clocks, happy faces, neutral faces, and sad faces), as well as the slope of change in the longest fixation duration (in milliseconds per trial) for repeated and novel stimuli across trials, reflecting the participants' attention patterns over time.

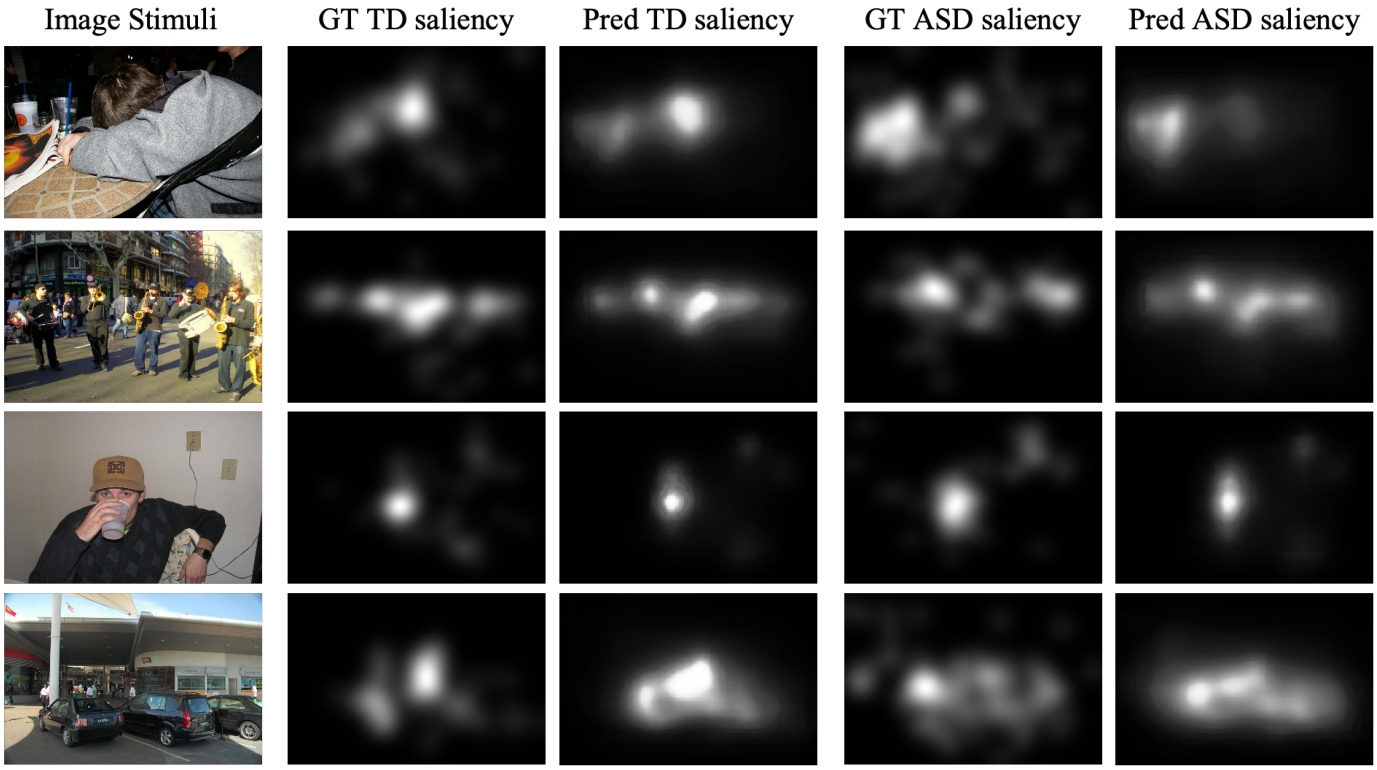

Fig. 2: Comparison of predicted and ground truth saliency maps for ASD and TD children on the Saliency4ASD dataset. Each row presents, from left to right: (1) the image stimulus, (2) the ground truth saliency map for TD children, (3) the predicted saliency map for TD children, (4) the ground truth saliency map for ASD children, and (5) the predicted saliency map for ASD children.

Fig. 3: Examples of the five stimulus categories used in the habituation paradigm: simple shapes, clocks, happy faces, neutral faces, and sad faces.

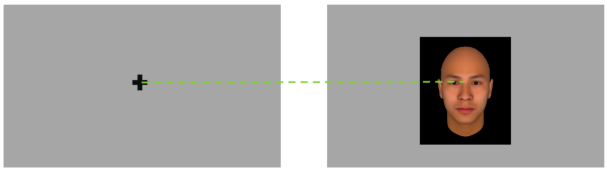

Fig. 4: Example of stimuli used in the vigilance/avoidance paradigm: a face where the eyes are centered.

B. Vigilance/Avoidance Paradigm

This paradigm investigates attention allocation to facial features, such as eyes and mouths, in controlled conditions. It consists of 16 trials, with half displaying faces where the mouth is centered and half displaying faces where the eyes are centered, as shown in Fig. 4. This figure illustrates an example of a face where the eyes are centered.

Each trial begins with a central fixation image to attract attention, followed by the appearance of a facial image aligned vertically with the previous fixation point. The face remains visible for 2.5 seconds, during which the participant's eye movements are recorded. The total experiment duration is 48 seconds.

The extracted features include the proportion of time spent fixating on eyes or mouth relative to the total face viewing time (`Proportion_eye` and `Proportion_mouth`), the proportion of second fixations directed to eyes, mouth, or face (`PropSecond_eye`, `PropSecond_mouth`, `PropSecond_face`), the latency to disengage from eyes or mouth (`eye_latency_data` for eye trials and `mouth_latency_data` for mouth trials), and the latency to look at the eyes after the mouth is initially centered (`eye_direction_latency`). For each participant, data from trials directed toward eyes and trials directed toward mouths are combined with prefixes `eyeDir_` and `mouthDir_` to distinguish conditions. Latency-related features are averaged across valid trials for each condition to calculate the final reported values.

Repetition Preference Paradigm

This paradigm evaluates participants' visual preferences for repetitive versus random motion stimuli. It comprises six trials. Each trial begins with a fixation image displayed centrally to attract attention, followed by two videos shown side by side: one depicting repetitive motion and the other random motion. Each trial lasts 30 seconds, resulting in a total duration of 3 minutes.

The extracted features include the Repetition Preference Index (RPI), which quantifies the participant's preference for repetitive motion based on gaze data, and the Randomness Preference Index (RanPI), which quantifies the participant's preference for random motion based on gaze data.

REFERENCES

[1] Jesús Gutiérrez, Zhaohui Che, Guangtao Zhai, and Patrick Le Callet, "Saliency4asd: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder," *Signal Processing: Image Communication*, vol. 92, pp. 116092, 2021.

[2] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu, "Transalnet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, 2022.

[3] Wei Ni, Haoyang Lu, Qiandong Wang, Ci Song, and Li Yi, "Vigilance or avoidance: How do autistic traits and social anxiety modulate attention to the eyes?," *Frontiers in Neuroscience*, vol. 16, pp. 1081769, 2023.

[4] Iti Arora, Alessio Bellato, Teodora Gliga, Danielle Ropar, Puja Kochhar, Chris Hollis, and Madeleine Groom, "What is the effect of stimulus complexity on attention to repeating and changing information in autism?," *Journal of Autism and Developmental Disorders*, pp. 1–17, 2021.

[5] Tianbi Li, Yewei Li, Yixiao Hu, Yuyin Wang, Cheuk Man Lam, Wei Ni, Xueqin Wang, and Li Yi, "Heterogeneity of visual preferences for biological and repetitive movements in children with autism spectrum disorder," *Autism Research*, vol. 14, no. 1, pp. 102–111, 2021.

(a) 30 images selected by our Image Selection Module, exhibiting the most significant eye gaze differences between ASD and TD children. These images serve as the primary stimuli for evaluating eye gaze behavior in ASD screening.

(b) 30 images showing minimal gaze pattern differences between ASD and TD children, serving as a control group for comparison.

Fig. 5: Image stimuli used in the study for collecting eye gaze data from participants: (a) images with significant eye gaze differences and (b) images with minimal eye gaze differences.