GIBRNet: A Multimodal Spatiotemporal Reasoning Network Integrating Emotion, Gaze, and Position for Gaze Interaction Behavior Recognition

Junhao Xiao[1], Jing Gao[1], Jingxing Zhong[2], Zhiyu Wu[3], Yi Chen[1†]
[1]School of Computer, Central China Normal University, Wuhan, Hubei, China
[2]Maynooth International Engineering College, Fuzhou University, Fuzhou, Fujian, China
[3]School of Computer Science, Fudan University, Shanghai, China
[†]Corresponding author: chenyi30@ccnu.edu.cn

Abstract

Gaze Interaction Behavior Recognition (GIBR) plays a significant role in understanding social behaviors and diagnosing mental health conditions. However, existing methods are limited by inadequate task modeling, resulting in suboptimal performance. To address this issue, we model the GIBR task as a spatiotemporal reasoning problem integrating three modalities: emotion, gaze, and position. Based on this, we propose GIBRNet, which enhances the representation of gaze interaction tendencies through an Emotion-Aware Refinement Matrix and dynamically aggregates multi-frame, multi-modal information using GP-GNN, enabling more precise interaction behavior reasoning. Comparative experiments on the VACATION dataset demonstrate that GIBRNet significantly outperforms existing approaches. Additionally, we constructed a GIBR dataset suite, consisting of three extended datasets, for generalization evaluation, demonstrating GIBRNet's superiority. All datasets and code are publicly available[1].

Keywords: Gaze Interaction Behavior Recognition; Multimodal Information Fusion; Social Behavior Understanding

Introduction

Gaze interaction is a fundamental mode of communication in social contexts. From a psychological perspective, the eyes are considered a cognitively special stimulus, with a "hard-wired" pathway in the brain dedicated to interpreting gaze behaviors (Vickers, 2007). Gaze interactions take various forms, each silently conveying the emotions and attitudes of participants (Loeb, 1972; Normoyle et al., 2013). By observing gaze interaction behaviors between individuals, we can infer their mental states, social intentions, and interpersonal relationships (Mutlu et al., 2012). For example, mutual gaze often signifies trust and intimacy, while gaze avoidance may indicate disinterest or shyness. Furthermore, atypical gaze behaviors are commonly used to diagnose mental disorders such as autism and schizophrenia (Franck et al., 2002; Pitskel et al., 2011). In educational settings, gaze interaction serves as an important indicator of students' collaboration fluency, enabling teachers to adjust groupings and enhance teaching efficiency (Whitehead et al., 2024). Therefore, developing methods for accurate Gaze Interaction Behavior Recognition (GIBR) contributes to a deeper understanding of interpersonal dynamics and plays a practical role in psychological diagnosis and educational optimization.

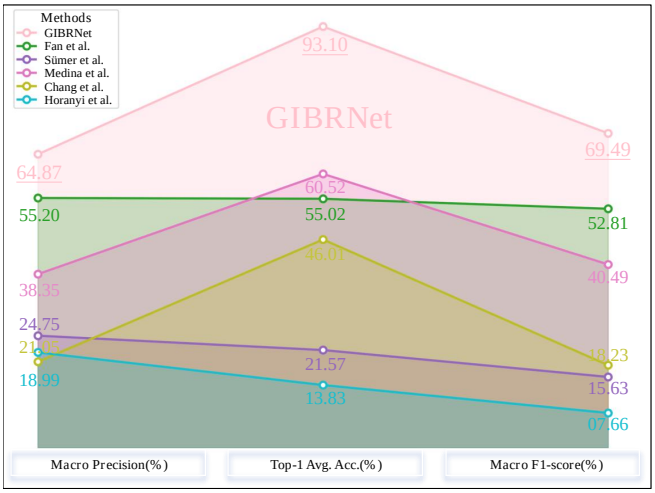

Figure 1: Performance comparison of GIBRNet with state-of-the-art methods on the VACATION dataset.

Cognitive psychology research (Cui et al., 2023; Loeb, 1972; Ottoni & Cerqueira, 2024) confirms that emotion, gaze, and position information are closely tied to social interactions. Although recent computer vision studies have advanced GIBR (Chang et al., 2023; Doosti et al., 2021; Fan et al., 2018a, 2019; Marin-Jimenez et al., 2019; Marin-Jimenez et al., 2014; Medina, 2021; Park et al., 2012; Soo Park & Shi, 2015; Sumer et al., 2020), suboptimal utilization of emotional, gaze, and positional cues, along with flawed task modeling, has hindered performance. We propose modeling GIBR as a spatiotemporal reasoning problem that integrates these multimodal signals. Based on this modeling, we introduce GIBRNet, which addresses the shortcomings of existing studies and significantly enhances recognition performance, as illustrated in Figure 1.

GIBRNet takes as input a sequence of video frames, the position information of interaction nodes, and a gaze adjacency matrix representing gaze directions. GIBRNet first extracts Node Position Features, Node Image Features, and Node Depth Features of interaction nodes from each frame using three convolutional feature extractors. The image and position features are concatenated into Node Emotion-Aware Features. GIBRNet leverages an attention mechanism (Vaswani, 2017) by treating the Node Emotion-Aware Features of each

[1]Datasets and code are publicly available at: https://github.com/Codecode-X/GIBRNet.git

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

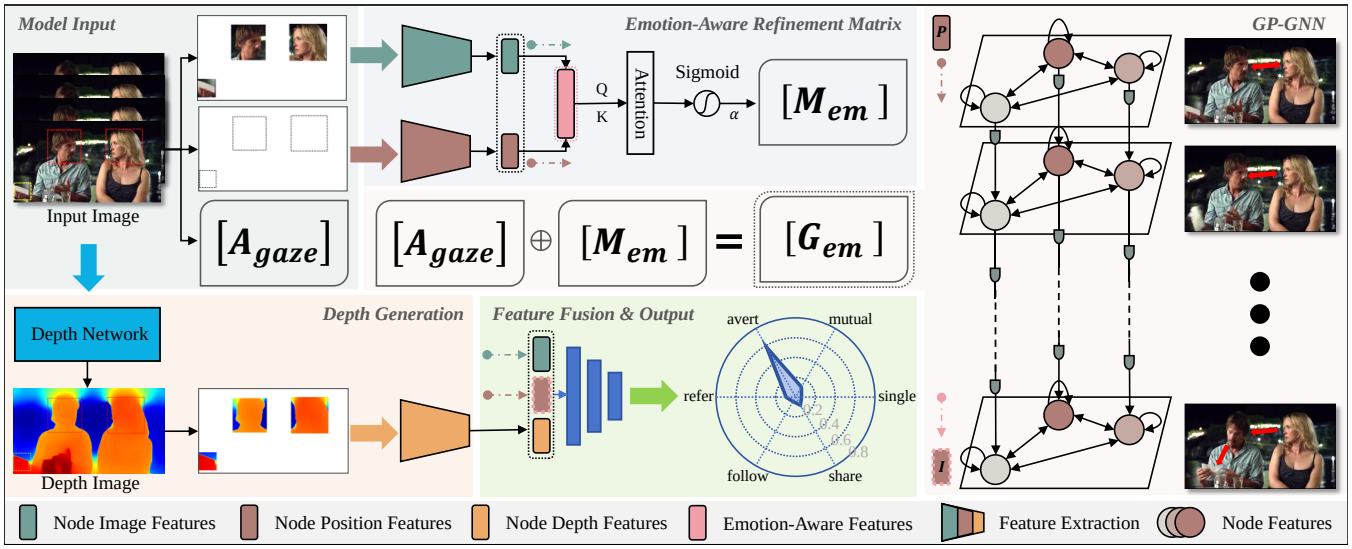

Figure 2: Flowchart of the GIBRNet: The Emotion-Aware Gaze Matrix $\mathbf{G}_{\text{em}}$ is obtained by fusing the Gaze Adjacency Matrix $\mathbf{A}_{\text{gaze}}$ with the Emotion-Aware Refinement Matrix $\mathbf{M}_{\text{em}}$. This matrix guides the spatial-temporal feature propagation of Node Position Features $\mathbf{P}$ in GP-GNN, generating Node Interaction Features $\mathbf{I}$. Finally, image, interaction, and depth features are fused for classification. $\mathbf{G}_{\text{em}}$ dynamically aggregates multi-frame, multi-modal information using GP-GNN, enabling more precise interaction behavior reasoning.

node pair as the Query and Key. It computes attention weights to form an Emotion-Aware Refinement Matrix, which integrates emotional cues to refine the Gaze Adjacency Matrix—representing direct gaze interactions—resulting in an Emotion-Aware Gaze Matrix that more accurately captures interaction tendencies between nodes.

To effectively integrate spatiotemporal interaction information among nodes, we propose GP-GNN. Guided by the Emotion-Aware Gaze Matrix, GP-GNN dynamically aggregates position information from neighboring nodes in the current frame and its own past state from the previous frame, generating Node Interaction Features that better capture complex dynamic interactions. Finally, the Node Image Features (representing emotion information), Node Interaction Features (integrating gaze and position information), and Node Depth Features (enhancing position information) are fused and fed into a classification head to infer the types of gaze interactions. This approach enables GIBRNet to effectively capture the temporal and spatial dimensions of emotion, gaze, and position information in videos, facilitating the accurate classification of complex gaze interaction patterns.

To comprehensively evaluate the performance and generalization ability of GIBRNet, we improved and extended existing datasets. This includes filtering incorrect annotations in the VACATION dataset (Fan et al., 2019) and augment UCO-LAEO, UCO-AVA (Marin-Jimenez et al., 2019), and VideoCoAtt (Fan et al., 2018b) with gaze relationships and co-attention object bounding boxes. The resulting three datasets, named G-UCO-LAEO, G-UCO-AVA, and G-VideoCoAtt, are hereafter collectively referred to as the GIBR datasets suite. Comparative experiments on the VACATION dataset demonstrate GIBRNet's significant advantage over existing methods. Generalization experiments on the GIBR dataset suite further confirm that GIBRNet also performs well in single gaze interaction recognition tasks.

Our contributions are as follows: **1)** We model the GIBR task as a spatiotemporal reasoning problem that integrates multimodal information, including emotion, gaze, and position, and propose the Multimodal Spatiotemporal Reasoning Network, GIBRNet. **2)** We improve and extend existing gaze interaction datasets by filtering errors in the VACATION dataset and constructing the GIBR datasets suite tailored for GIBR tasks. **3)** Experiments show GIBRNet significantly outperforms state-of-the-art methods in GIBR tasks.

Related Work

In recent years, research on GIBR has attracted considerable attention due to its wide range of application values in social behavior understanding, mental health assessment, and educational scenario optimization.

Joint Attention Recognition. Park et al. (2012) and Soo Park and Shi (2015) proposed a concept of social saliency based on a first-person perspective. By leveraging head-mounted camera videos, they inferred group-level joint attention patterns, pioneering a new direction in social scene analysis. Fan et al. (2018a) introduced a method to infer joint attention in a third-person perspective. They employed a gaze estimation and region proposal module to extract features and used spatial detection and temporal optimization modules to identify and refine the joint attention heatmap. Sumer et al. (2020) proposed an end-to-end approach that combines saliency and joint attention. By encoding the positions of

faces and scenes in dual-channel heatmaps, they effectively improved joint attention recognition accuracy.

Mutual Gaze Detection. Marin-Jimenez et al. (2014) investigated mutual gaze patterns by modeling head angles with Gaussian process regression, integrating relative positions to infer a "mutual gaze" score. Doosti et al. Doosti et al. (2021) enhanced performance by incorporating an auxiliary 3D gaze estimation task, thereby refining head feature representation without increasing annotation costs. Marin-Jimenez et al. (2019) and Medina (2021) extended their research to video trajectories, proposing a head pose branch with 3D convolutions and a spatiotemporal feature fusion method, which further improved mutual gaze detection accuracy.

Gaze Behavior Interaction Recognition. Fan et al. (2019) defined six types of gaze interaction patterns and developed a spatiotemporal graph neural network along with an encoder-decoder event network to infer gaze interaction types and events among individuals. Chang et al. (2023) categorized static gaze interactions into five types, using ResNet to extract features and generate head location embeddings, which effectively detected gaze interaction states.

Although these studies have advanced GIBR, existing methods lack effective modeling, underutilize key multimodal information, and fail to precisely process spatiotemporal correlations among nodes.

GIBR Task Definition and Modeling

Definition

Gaze Interaction Behavior Recognitio (GIBR) lies at the intersection of computer vision and cognitive psychology, aiming to uncover interaction patterns via gaze behavior analysis in videos. The task takes as input video frames, node positions (identified by bounding boxes), and a gaze adjacency matrix. Nodes include both individuals and objects involved in interactions.

Based on these inputs, GIBR requires classifying the gaze interaction behavior between the two persons in the video into six types: *Single, Mutual, Avert, Refer, Follow*, and *Share*.

Modeling

According to research in cognitive psychology, emotion, gaze, and position information are all closely related to interpersonal interactions. Emotional states influence gaze behavior (Ottoni & Cerqueira, 2024)—positive emotions often increase eye contact, whereas negative emotions may lead to gaze avoidance. Gaze dynamics reflect social relationships and intentions (Loeb, 1972), while personal space theory (Cui et al., 2023) emphasizes that position information is critical for interpreting social interaction, indicating that individuals' spatial usage is directly linked to their modes of social engagement.

We model GIBR as a spatiotemporal reasoning problem based on the multimodal fusion of emotion, gaze, and position features. This approach exploits these three modalities across spatial and temporal dimensions to learn a function f that optimally predicts interaction outcomes.

Given N interacting entities in a scene, each entity at time t is represented by a feature vector $\mathbf{x}_t^i = \left[\mathbf{e}_t^i,\ \mathbf{g}_t^i,\ \mathbf{p}_t^i\right]$, where $\mathbf{e}_t^i \in \mathbb{R}^d$ denotes emotion features, $\mathbf{g}_t^i \in \mathbb{R}^3$ represents gaze features, and $\mathbf{p}_t^i \in \mathbb{R}^3$ encodes position features. Collecting these features over T time steps forms $\mathbf{X} = \{\mathbf{x}_t^i \mid t = 1,\ldots,T,\ i = 1,\ldots,N\}$. The goal is to learn a spatiotemporal reasoning function $f : \mathbf{X} \mapsto \mathbf{y}$, where $\mathbf{y}$ represents the predicted interaction behavior.

By jointly considering emotion, gaze, and position in both spatial and temporal dimensions, this modeling approach enables more accurate recognition of complex gaze interaction behaviors among nodes.

Our Approach

Based on our modeling of the GIBR task, we propose GIBRNet, as illustrated in Figure 2, designed to fully leverage spatiotemporal multimodal information in video sequences for the analysis and reasoning of complex gaze interaction behaviors. The design of GIBRNet consists of multiple stages, including feature extraction, emotion-aware refinement, spatiotemporal information aggregation, and multimodal fusion. Each stage addresses different aspects of interaction behavior, ultimately enabling the model to accurately capture patterns of social interaction.

Feature Extraction

The feature extraction in GIBRNet is performed by three convolutional neural network (CNN)-based modules: position, image, and depth feature extractors. The position feature extractor derives the Node Position Features $\mathbf{P}_{t,n}$ from the bounding box information of each node, capturing the spatial positioning to reflect the physical distances and configurations among nodes. The image feature extractor extracts Node Image Features $\mathbf{X}_{t,n}$ from the cropped image regions of the node's bounding box, capturing the visual appearance and implicit emotion information of nodes for analyzing emotional communication. The depth feature extractor computes Node Depth Features $\mathbf{D}_{t,n}$ from the depth maps generated by the MiDas (Ranftl et al., 2020), providing three-dimensional position information to enhance scene perception.

Emotion-Aware Refinement Matrix

Relying solely on gaze information may not sufficiently reflect interaction tendencies among nodes, as emotional cues should also be considered (Micheli et al., 2024; Ottoni & Cerqueira, 2024). To address this, we compute an Emotion-Aware Refinement Matrix to refine the direct gaze information, resulting in an Emotion-Aware Gaze Matrix that better captures interaction tendencies among nodes.

First, emotions are inferred through facial expressions and interpersonal distances, as prior studies have emphasized the need to jointly consider these two factors for reliable emotion inference (Hsu et al., 2024; Lebert et al., 2024). We extract facial expression information from the Node Image Fea-

tures $\mathbf{X}_{t,n}$ and interpersonal distances from the Node Position Features $\mathbf{P}_{t,n}$. These are concatenated to form the Emotion-Aware Features $\mathbf{E}_{t,n} = \mathbf{X}_{t,n} || \mathbf{P}_{t,n}$.

Next, for each pair of nodes, their Emotion-Aware Features $\mathbf{E}_{t,n}$ are used as query (Q) and key (K) inputs to an attention model (Vaswani, 2017). Specifically, node i's features $\mathbf{E}_{t,i}$ serve as the query, while node j's features $\mathbf{E}_{t,j}$ act as the key. The attention score is computed as $\alpha_{tij} = \sigma(\mathbf{E}_{t,i}\mathbf{E}_{t,j}^T)$, where σ is the sigmoid activation function that maps the dot product to the range $(0,1)$. The attention scores for all node pairs populate the Emotion-Aware Refinement Matrix $\mathbf{M}_{\text{em},t}$, which reflects interaction tendencies based on emotion information.

Finally, the Emotion-Aware Refinement Matrix $\mathbf{M}_{\text{em},t}$ is used to refine the direct gaze information $\mathbf{A}_{\text{gaze},t}$, yielding the Emotion-Aware Gaze Matrix $\mathbf{G}_{\text{em},t} = \mathbf{A}_{\text{gaze},t} + \mathbf{M}_{\text{em},t}$.

The Emotion-Aware Gaze Matrix integrates emotion and gaze information, enabling more accurate capture of interaction tendencies among nodes and providing precise guidance for subsequent information transmission between nodes.

GP-GNN

To better integrate spatiotemporal interaction information between nodes and understand complex dynamic interactions, we propose GP-GNN (GazePosition-GNN), a spatiotemporal graph neural network (Gori et al., 2005; Kipf & Welling, 2016; Li et al., 2024), specifically designed to compute the node interaction feature $\mathbf{I}_{t,n}$ for each node in each frame.

In GP-GNN, the interaction feature $\mathbf{I}_{t,n}$ of each node is calculated by aggregating the Node Position Features $\mathbf{p}_{t,m}$ of neighboring nodes within the same frame, guided by the Emotion-Aware Gaze Matrix $\mathbf{G}_{\text{em},t}$, and incorporating the node's interaction features $\mathbf{I}_{t-1,n}$ from the previous frame. The aggregation method is defined as:

$$\mathbf{I}_{t,n} = \sum_m \mathbf{G}_{\text{em},t}[n,m] \cdot \mathbf{p}_{t,m} + \mathbb{I}_{t>0} \cdot \mathbf{f}(\mathbf{I}_{t-1,n}),$$

where $\mathbf{G}_{\text{em},t}[n,m]$ denotes the information weight from node n to node m, reflecting interaction tendencies based on gaze and emotion information. $\mathbb{I}_{t>0}$ is an indicator function ensuring that features from the previous frame are considered only for non-initial frames. $\mathbf{f}(\cdot)$ is an adaptation function implemented via a convolutional layer to adjust and propagate interaction features from the previous timestep, capturing temporal dependencies.

By employing this multimodal spatiotemporal aggregation strategy, GP-GNN effectively captures complex gaze interaction behavior patterns.

Multimodal Information Fusion and Classification

We first concatenate the multimodal features of each node, including image features $\mathbf{X}_{t,n}$ (capturing emotion information), depth features $\mathbf{D}_{t,n}$ (enhancing position information), and interaction features $\mathbf{I}_{t,n}$ (integrating gaze and position information), along the feature dimension to obtain node-level features as $\mathbf{F}_{t,n} = \mathbf{X}_{t,n} || \mathbf{D}_{t,n} || \mathbf{I}_{t,n}$. Next, the features of all nodes within the same frame are further concatenated to form frame-level features $\mathbf{F}_t$, which are then flattened across all frames of the video segment to produce video-level features $\mathbf{F}_{\text{video}} = \text{Flatten}(\{\mathbf{F}_t\}_{t=1}^T)$. Based on this, we feed $\mathbf{F}_{\text{video}}$ into a classification head composed of fully connected layers and ReLU activation functions alternately. The probability distribution of each social interaction category is computed using the Softmax function (Bridle, 1989) as $\mathbf{P}_{\text{class}} = \text{Softmax}(\mathbf{h}(\mathbf{F}_{\text{video}}))$, where $\mathbf{h}$ denotes nested fully connected layers and ReLU operations (Nair & Hinton, 2010).

By integrating emotion, gaze, and position modalities, GIBRNet effectively captures implicit emotional expressions and spatial configuration features within gaze interactions. This enables a more nuanced interpretation of complex behavioral patterns, ultimately improving gaze interaction behavior category prediction.

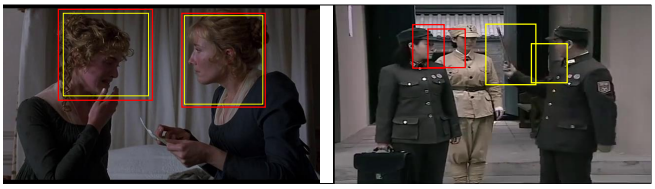

Figure 3: Left shows misaligned head and gaze target labels; right wrongly labels right person as left's gaze target.

Training and Testing

Since GIBRNet is designed for a multi-class classification problem, we adopt the cross-entropy loss (Mao et al., 2023) during training. An SGD optimizer (Amari, 1993) is used, and training proceeds in three steps on the training set with learning rates 0.001, 0.0005, and 0.0001, respectively. After training, we save the model parameters from the best-performing step and evaluate on the test set by analyzing the confusion matrix of classification results. To mitigate overfitting, the weight decay is set to 0.0001.

Dataset Preparation

To validate our model, we prepared multiple datasets for comparative experiments and generalization tests.

VACATION Dataset

VACATION (Fan et al., 2019) is an open-source gaze interaction dataset with 300 TV show clips totaling 96,993 frames. It features diverse social scenes, cultural backgrounds, lighting, resolutions, and occlusions, making it an ideal benchmark for complex gaze interaction recognition. However, during experimentation, we found some annotation errors in VACATION, as illustrated in Figure 3, which could compromise the validity of the results. Consequently, we discarded the erroneous samples.

GIBR Datasets Suite

Beyond VACATION, datasets like UCO-LAEO, AVA-LAEO (Marin-Jimenez et al., 2019), and VideoCoAtt (Fan et

Method	Macro F1	Macro Prec.	Top-1 Avg. Acc.
Ours full	**69.495**	**64.87**	**93.1**
w/o Depth	63.29	59.48	89.57
w/o $\mathbf{M}_{\text{em}}$	56.4	54.7	74.18
$\mathbf{E} \rightarrow \mathbf{I}$	62.86	59.3	89.7

Table 1: Macro-level results of ablation experiments.

Method	Single		Mutual	
	F1	Prec.	F1	Prec.
Ours full	**92.82**	98.27	**99.69**	**99.39**
w/o Depth	87.72	**98.60**	99.28	99.39
w/o $\mathbf{M}_{\text{em}}$	83.27	94.63	97.37	99.08
$\mathbf{E} \rightarrow \mathbf{I}$	82.98	98.51	99.69	99.39
Method	**Avert**		**Refer**	
	F1	Prec.	F1	Prec.
Ours full	6.62	3.82	**66.67**	**55.81**
w/o Depth	**11.29**	**6.02**	53.49	40.35
w/o $\mathbf{M}_{\text{em}}$	3.57	2.03	44.90	31.88
$\mathbf{E} \rightarrow \mathbf{I}$	11.11	5.92	63.01	52.27
Method	**Follow**		**Share**	
	F1	Prec.	F1	Prec.
Ours full	**56.96**	**41.28**	94.21	90.65
w/o Depth	32.73	19.91	**95.25**	**92.60**
w/o $\mathbf{M}_{\text{em}}$	16.36	9.00	92.95	91.58
$\mathbf{E} \rightarrow \mathbf{I}$	32.03	19.40	88.32	80.29

Table 2: Micro-level results of ablation experiments.

al., 2018b) focus on single gaze interaction behaviors. UCO-LAEO includes 129 TV show clips with head bounding boxes and mutual gaze labels. AVA-LAEO adds binary mutual gaze labels to the AVA dataset. VideoCoAtt, derived from 20 YouTube TV shows, features 380 sequences and 492,100 frames, annotated for shared attention.

To accommodate the GIBR task, we curated and converted UCO-LAEO and UCO-AVA by adding annotations for gaze relationships among nodes, thus creating G-UCO-LAEO and G-UCO-AVA. Similarly, we augmented VideoCoAtt with node-level gaze annotations and bounding boxes for jointly attended objects, yielding G-VideoCoAtt. These additional datasets form the GIBR dataset suite, which enriches testing scenarios and enables more comprehensive evaluations of our method's generalization capabilities.

Experiments

Metrics

To comprehensively evaluate our GIBRNet model for gaze interaction behavior recognition, we employ the following metrics. **Precision** is defined as the proportion of positive predictions that are truly positive, calculated as $\frac{\text{TP}}{\text{TP + FP}}$, where TP represents true positives and FP false positives. **Recall** measures the proportion of actual positive samples correctly predicted, expressed as $\frac{\text{TP}}{\text{TP + FN}}$, where FN represents false negatives. The **F1-Score**, calculated as $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, provides a balance between Precision and Recall. **Macro Precision** and **Macro F1-Score**, as arithmetic means of Precision and F1-Scores across all classes, provide overall measures unaffected by class imbalance. Lastly, **Top-1 Avg.Acc.**, the proportion of correctly predicted samples over the entire test set, offers a straightforward evaluation of overall accuracy.

Method	Macro F1	Macro Prec.	Top-1 Avg. Acc.
CNN (Ours)	**69.49**	**64.87**	**93.10**
Fan et al.	52.81	55.20	55.02
Sümer et al.	15.63	24.75	21.57
Medina et al.	40.49	38.35	60.52
Chang et al.	18.23	18.99	46.01
Horanyi et al.	7.66	21.05	13.83

Table 3: Macro-level results of comparative experiments.

Ablation Study

To validate the independent contribution of each module or the optimality of strategies in GIBRNet, we conducted ablation experiments by controlling specific module variables. **Ours full** represents the complete GIBRNet structure. **w/o $\mathbf{M}_{\text{em}}$** excludes the Emotion-Aware Refine Matrix $\mathbf{M}_{\text{em}}$ used to refine direct gaze information $\mathbf{A}_l$. **w/o Depth** omits monocular depth information of nodes. $\mathbf{E} \rightarrow \mathbf{I}$ indicates that the information propagated between nodes in GP-GNN is Emotion-Aware Features rather than Node Position Features.

The experimental results in Tables 1 and 2 demonstrate, from both macro and micro perspectives, the independent contribution of each module and the optimality of strategies in GIBRNet:

w/o Depth. Adding depth information improves all macro-level metrics (Macro F1-Score, Macro Precision, Top-1 Avg.Acc.) and boosts recognition precision and accuracy in *single*, *mutual*, *refer*, and *follow*, enhancing GIBRNet's spatial understanding and ability to capture inter-node distances. However, performance drops slightly in *avert* and *share*, possibly due to over-reliance on depth for interpreting complex social dynamics. Future improvements should balance depth with other modalities.

w/o $\mathbf{M}_{\text{em}}$. Incorporating the Emotion-Aware Refine Matrix $\mathbf{M}_{\text{em}}$ to refine the gaze adjacency matrix $\mathbf{A}_{\text{gaze}}$ improves all macro- and micro-level metrics, showing that gaze alone cannot fully capture interaction tendencies and that emotional cues are essential.

$\mathbf{E} \rightarrow \mathbf{I}$. Propagating and updating only position features in GP-GNN improves performance by 3%, showing position information dominates this task, while image features in emotional cues may dilute its influence.

Method	Single		Mutual		Avert		Refer		Follow		Share	
	F1	Prec.	F1	Prec.	F1	Prec.	F1	Prec.	F1	Prec.	F1	Prec.
Ours	**92.82**	**98.27**	**99.69**	**99.39**	6.62	3.82	**66.67**	55.81	**56.96**	**41.28**	**94.21**	**90.65**
Fan et al.	26.17	22.10	98.60	98.68	**74.28**	**59.20**	53.16	**56.90**	18.05	32.83	46.61	61.51
Sümer et al.	28.04	46.01	32.88	46.57	0.00	0.00	2.51	1.28	3.68	1.90	26.68	52.74
Medina et al.	61.19	79.56	80.03	74.75	5.45	2.95	17.24	11.49	32.58	20.93	46.47	40.41
Chang et al.	58.71	51.84	38.64	38.46	0.00	0.00	0.00	0.00	0.00	0.00	12.05	23.65
Horanyi et al.	27.15	35.71	14.22	41.45	1.10	0.55	1.81	0.97	0.00	0.00	1.66	47.62

Table 4: Micro-level results of comparative experiments.

Dataset	Total Samples	Detected Samples for Each Type						Precision/Recall(%)
		Single	Mutual	Avert	Refer	Follow	Share	
G-UCO-LAEO (Mutual)	100	0	**100**	0	0	0	0	100
G-AVA-LAEO (Mutual)	100	0	**99**	1	0	0	0	99
G-VideoCoAtt (Share)	200	0	0	0	0	0	**200**	100

Table 5: Results of generalization experiments.

Real Type	single	mutual	avert	refer	follow	share	Recall
single	**2438**	9	126	19	60	120	87.95
mutual	0	**1468**	0	0	0	0	100
avert	15	0	**5**	0	0	0	25
refer	1	0	0	**24**	4	0	82.76
follow	4	0	0	0	**45**	0	91.84
share	23	0	0	0	0	**1164**	98.06

Table 6: Confusion matrix and per-class recall (%) from the comparative experiments.

Comparative Experiments

To validate the performance of GIBRNet, we compare GIBRNet with existing methods, with results in Tables 3 and 4.

At the macro level, GIBRNet achieves significantly higher Macro F1-Score, Macro Precision, and Top-1 Avg.Acc. than baselines, surpassing (Fan et al., 2019) by 33.19% and 17.52% on Macro F1-Score and Macro Precision, respectively, and (Medina, 2021) by 53.84% on Top-1 Avg.Acc.

At the micro level, GIBRNet achieves the highest F1-scores on *single*, *mutual*, *refer*, *follow*, and *share*, and the highest Precision on *single*, *mutual*, *follow*, and *share*. Precision and F1-scores surpass 90% for *single*, *mutual*, and *share*.

However, the *avert*, *refer*, and *follow* classes exhibit lower Precision and F1-scores. Table 6's confusion matrix suggests that this likely due to their significantly smaller sample sizes.

Overall, these results confirm that GIBRNet—which comprehensively considers emotion, gaze, and position across temporal and spatial dimensions—exhibits effectiveness and state-of-the-art performance for the GIBR task.

Generalization Experiments

To further assess the generalization capability of our method, we evaluate GIBRNet on the GIBR dataset suite, which we curated and annotated to suit the GIBR task. These tests focus on single-category gaze interaction recognition, specifically *mutual* for G-UCO-LAEO and G-UCO-AVA, and *share* for G-VideoCoAtt. The results are summarized in Table 5.

The experiments demonstrate that GIBRNet maintains robust generalization on detecting *mutual* gaze and *share* interactions, underscoring its strong adaptability. These outcomes confirm the superior generalizability of GIBRNet for single-category gaze interaction recognition.

Conclusion

In this paper, we propose GIBRNet, a spatiotemporal reasoning model integrating emotion, gaze, and position information for efficient gaze interaction behavior recognition. We validate its superior performance over existing methods on the VACATION dataset and further demonstrate its strong generalization capability on the newly constructed GIBR datasets suite. In future work, we plan to explore additional modalities (e.g., audio signals) to further enhance the model's understanding of complex social interactions.

Acknowledgments

This research was supported by the General Program of the National Natural Science Foundation of China (No. 62377024), titled *Discovering Interpretable Cognitive Evolution Paths Based on Causal Reasoning*, running from January 1, 2024 to December 31, 2027.

References

Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, *5*(4-5), 185–196.

Bridle, J. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in neural information processing systems*, *2*.

Chang, F., Zeng, J., Liu, Q., & Shan, S. (2023). Gaze pattern recognition in dyadic communication. *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, 1–7.

Cui, S., Yang, T., & Liu, N. (2023). Generalization of the modulatory effect of social interaction on personal space. *Frontiers in Psychology*.

Doosti, B., Chen, C.-H., Vemulapalli, R., Jia, X., Zhu, Y., & Green, B. (2021). Boosting image-based mutual gaze detection using pseudo 3d gaze. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(2), 1273–1281.

Fan, L., Chen, Y., Wei, P., Wang, W., & Zhu, S.-C. (2018a). Inferring shared attention in social scene videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6460–6468.

Fan, L., Chen, Y., Wei, P., Wang, W., & Zhu, S.-C. (2018b). Inferring shared attention in social scene videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6460–6468.

Fan, L., Wang, W., Huang, S., Tang, X., & Zhu, S.-C. (2019). Understanding human gaze communication by spatio-temporal graph reasoning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5724–5733.

Franck, N., Montoute, T., Labruyère, N., Tiberghien, G., Marie-Cardine, M., Daléry, J., d'Amato, T., & Georgieff, N. (2002). Gaze direction determination in schizophrenia. *Schizophrenia research*, *56*(3), 225–234.

Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.*, *2*, 729–734.

Hsu, C.-T., Sato, W., & Yoshikawa, S. (2024). An investigation of the modulatory effects of empathic and autistic traits on emotional and facial motor responses during live social interactions. *Plos one*, *19*(1), e0290765.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lebert, A., Vergilino-Perez, D., & Chaby, L. (2024). Keeping distance or getting closer: How others' emotions shape approach-avoidance postural behaviors and preferred interpersonal distance. *Plos one*, *19*(2), e0298069.

Li, Z., Xia, L., Xu, Y., & Huang, C. (2024). Gpt-st: Generative pre-training of spatio-temporal graph neural networks. *Advances in Neural Information Processing Systems*, *36*.

Loeb, B. K. (1972). Mutual eye contact and social interaction and their relationship to affiliation.

Mao, A., Mohri, M., & Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications. *International conference on Machine learning*, 23803–23828.

Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., & Zisserman, A. (2019). Laeo-net: Revisiting people looking at each other in videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3477–3485.

Marin-Jimenez, M. J., Zisserman, A., Eichner, M., & Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, *106*, 282–296.

Medina, M.-J. M. K. V. (2021). Suárez p zisserman a laeo-net++: Revisiting people looking at each other in videos. *IEEE Trans. Pattern Anal. Mach. Intell*, *10*.

Micheli, L., Breil, C., & Böckler, A. (2024). Golden gazes: Ngaze direction and emotional context promote prosocial behavior by increasing attributions of empathy and perspective-taking. *Journal of Personality and Social Psychology*, *126*(4), 643.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *1*(2), 1–33.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814.

Normoyle, A., Badler, J. B., Fan, T., Badler, N. I., Cassol, V. J., & Musse, S. R. (2013). Evaluating perceived trust from procedurally animated gaze. In *Proceedings of motion on games* (pp. 141–148).

Ottoni, L. T. C., & Cerqueira, J. d. J. F. (2024). A systematic review of human–robot interaction: The use of emotions and the evaluation of their performance. *International Journal of Social Robotics*, 1–20.

Park, H. S., Jain, E., & Sheikh, Y. (2012). 3d gaze concurrences from head-mounted cameras. *NIPS*.

Pitskel, N. B., Bolling, D. Z., Hudac, C. M., Lantz, S. D., Minshew, N. J., Vander Wyk, B. C., & Pelphrey, K. A. (2011). Brain mechanisms for processing direct and averted gaze in individuals with autism. *Journal of autism and developmental disorders*, *41*, 1686–1693.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, *44*(3), 1623–1637.

Soo Park, H., & Shi, J. (2015). Social saliency prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4777–4785.

Sumer, O., Gerjets, P., Trautwein, U., & Kasneci, E. (2020). Attention flow: End-to-end joint attention estimation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3327–3336.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vickers, J. N. (2007). *Perception, cognition, and decision training: The quiet eye in action*. Human Kinetics.

Whitehead, R., Nguyen, A., & Järvelä, S. (2024). Exploring the role of gaze behaviour in socially shared regulation of collaborative learning in a group task. *Journal of Computer Assisted Learning*.