Heterogeneous Multi-Modal Multi-Label Behavioral Context Recognition Using Wearable Sensors and Mobile Devices

Haodong Liu, Ye Zhang, Zhidan Liu, *Member, IEEE,* Hongyuan Zhu, *Member, IEEE,* and Changhong Wang

***Abstract*—**Accurate behavioral context recognition is essential for advancing social intelligence in smart home settings and social companion pairing. With advancements in mobile computing technologies, wearable sensors and mobile devices have become crucial for multi-modal data collection to interpret behavioral context. However, existing methods may overlook the multi-label attribute of behavioral context, limiting their ability to model modality-to-label and label-to-label dependencies. Moreover, the unique label heterogeneity in behavioral context recognition, arising from category-specific differences among labels, is often underestimated. To overcome these challenges, we propose a heterogeneous multi-modal multi-label recognition method to capture the relationships between modalities and labels, while taking the label heterogeneity into account. Using specialized heterogeneous decoders with self-adaptive attention, and a heterogeneous graph attention network based on the dual-level attention mechanism, our method captures complex dependencies among labels and modalities. Experiments on two benchmark datasets demonstrate the superior performance of our approach, with ablation studies validating the contributions of each component.

***Index Terms*—Behavioral context recognition, Multi-modal learning, Multi-label learning**

I. Introduction

THE behavioral context of a person encompasses their environment, activities, and social interactions at any given moment, addressing questions such as: “where is the person?”, “what is the person doing?”, “who is with the person?” [1], [2]. Recognizing behavioral context is a fundamental task in ubiquitous and mobile computing systems, providing essential data support for numerous downstream applications and services. For instance, it supports the customization of smart home environments to align with individual behavior patterns [3], [4], facilitates accurate social context-aware companion pairing [5], [6], and provides the basis data support for analyzing mental health trends among members of society [7], [8], [9]. Therefore, accurate behavioral context recognition is crucial for enhancing social intelligence by advancing real-time human-computer interaction services and long-term behavioral modeling.

The advancement of mobile computing technologies has enabled wearable and mobile devices, such as smartwatches and smartphones, to incorporate a wide array of embedded sensors for capturing diverse modalities, including kinematic data, physiological signals, and environmental acoustics. These devices facilitate the unobtrusive and energy-efficient collection of personalized multi-modal data [10], thereby providing a comprehensive depiction of the user's behavioral context. Previous studies have introduced various machine learning methods in multi-modal behavioral context recognition [2], [11], with a particular emphasis on deep learning technologies [12], [13]. However, these methods normally underestimate the multi-label attribute of behavioral contexts, where a single instance can be characterized by multiple co-existing labels. For instance, a behavioral context such as “a person is strolling outside on the street while talking with friends” can be represented by the multi-label tuple: < “strolling”, “outside”, “on the street”, “talking”, “with friends” >. These labels describe different aspects of the same situation and often exhibit co-occurrence relationships. Ignoring such relationships prevents the model from exploiting the potential interactions, leading to reduced recognition performance.

The multi-label attribute of behavioral context introduces the intra labels connections itself and the inter modality-label connections with the multi-modal attribute. On the one hand, we observe that features from different modalities contribute differently to each behavioral context label, which can be

Haodong Liu and Changhong Wang are with School of Biomedical Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, Guangdong 518107 China. E-mail: liuhd26@alumni.sysu.edu.cn, wangchh55@mail.sysu.edu.cn.

Ye Zhang is with School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, Guangdong 518107 China. E-mail: zhangy2658@mail.sysu.edu.cn.

Zhidan Liu is with Intelligent Transportation Thrust, System Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511458 China. E-mail: zhidanliu@hkust-gz.edu.cn.

Hongyuan Zhu is with the Institute for Infocomm Research (I2R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore. E-mail: hongyuanzhu.cn@gmail.com

This work was supported in part by National Natural Science Foundation of China (Grant No: 62303496, 62573441), Shenzhen Medical Research Fund (Grant No: D250403003) and the Guangdong Basic and Applied Basic Research Foundation (Grant No: 2025A1515011729). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsor.

(Corresponding authors: Changhong Wang.)

defined as modality-to-label dependencies. For example, vigorous motion signals may indicate running activity, whereas stable audio signals might reflect a quiet environment such as a library or classroom. On the other hand, we observe that relevant labels tend to co-occur more frequently than conflicting labels, a relationship we define as label-to-label dependencies. For example, it is highly probable that a person would be sitting inside a car rather than standing inside it. Based on these observations, we hypothesize that simultaneously modeling modality-to-label and label-to-label dependencies could facilitate the interactions between modalities and labels, thereby improving the model's recognition ability.

Furthermore, behavioral context labels exhibit inherent heterogeneity, as they can be divided into distinct categories, each describing different aspects of behavioral contexts with unique semantic structures. This heterogeneity cause labels within the same category to share similar learning patterns, while labels across categories exhibit distinct patterns in modeling modality-to-label and label-to-label dependencies. For example, within the "phone position" category, labels like "phone in hand" and "phone in pocket" both describe the phone being carried on the limb or body, which reflect the user's physical interaction with the phone, thereby linking strongly to the motion modality. In contrast, labels in the "places" category, such as "in the library" and "in the classroom", are more related to environmental factors and thus strongly tied to the acoustic modality. These examples reflect how heterogeneity influences the modeling for modality-to-label dependencies. For label-to-label dependencies, labels within category "complex activities", such as "shopping" and "cooking", often involve specific environment, making them closely relate to labels in the "places" category. Conversely, labels in the "phone position" category do not heavily depend on specific locations, and thus are not strongly associated with the labels in the "places" category. Therefore, recognizing label heterogeneity enables the design of refined structures for capturing these dependencies, leading to enhanced recognition performance.

To leverage the above data characteristics, we propose a heterogeneous multi-modal multi-label behavioral context recognition approach that simultaneously models both modality-to-label and label-to-label dependencies, while accounting for label heterogeneity. This work is inspired by recent advances in multi-modal multi-label classification using Transformer and Graph Neural Network (GNN) [14], [15]. Building on these techniques, we further incorporate the heterogeneity among behavioral context labels into the network designing. The key contributions of our work are as follows:

1. To the best of our knowledge, we are the first to simultaneously notice both multi-modal and multi-label attributes in behavioral context recognition. To model modality-to-label dependencies, we introduce specialized decoders of self-adaptive attention, controlling each modality's contribution to labels. To model label-to-label dependencies, we adopt a graph attention-based neural network to capture the complex correlations among behavioral context labels.
2. To leverage the label heterogeneity for more effective dependency modeling, we adopt distinct learning strategies for each label category based on its unique semantic structure. By utilizing heterogenous decoder structures and a dual-level attention mechanism, we ensure consistent learning patterns for labels within the same category while enabling differentiation of learning patterns across distinct categories.
3. Comprehensive experiments validate the effectiveness of our approach in behavioral context recognition, underscoring its potential for future research. We also demonstrate the practical benefits of accurate context recognition in inferring mental health status, highlighting its real-world applicability. Our project is available at: https://github.com/Estampie00/HMMBCR

II. Related Works

A. Multi-modal Behavioral Context Recognition

Recognition of behavioral context using multi-modal sensors has been extensively explored [16]. Early research leveraged hand-crafted features combined with machine learning techniques for multi-modal behavioral context classification. Vaziman et al. integrated features separately extracted from the motion, audio, and phone state modalities, and then employed a Multilayer Perceptron (MLP) for classification [2]. Ehatisham-ul-Haq and Azam developed a dual-level classification methodology, initially recognizes behavioral labels to aid in the recognition of context labels [11].

The advent of deep learning has revolutionized the field, introducing powerful techniques such as Convolutional Neural Networks (CNNs). For instance, Saeed et al. proposed a multi-stream convolutional network to predict behavioral context labels in an end-to-end manner [12]. With the increasing prevalence of attention mechanisms, Bhattacharya et al. utilized a self-attention network to derive a cohesive representation of multi-modal behavioral context data [17]. Similarly, Yang et al. introduced a multi-scale cross-modal interactive network, performing feature-level fusion across various modalities [13]. However, these approaches either treat behavioral context recognition as a multi-class classification problem, neglecting the fact that real-world contexts often require multiple labels to describe, or they use separate binary classifiers for each label under the multi-label setting, failing to capture the complex dependencies between labels.

B. Multi-label Classification

Multi-label classification tasks seek to address the issue where each instance can simultaneously be associated with multiple labels. This approach is particularly useful in various fields such as text categorization [18], image recognition [19] , and bioinformatics [20], where complex data often exhibit multiple relevant labels simultaneously.

To capture label relationships, earlier researchers have attempted to model label dependencies by sequentially considering label co-occurrence. For example, Wang et al. proposed a CNN-RNN framework, which utilized a recurrent neural network (RNN) to model the co-occurrence label dependencies [21]. Similarly, Zhang et al. employed a CNN to extract semantic features and a long short-term memory (LSTM) to sequentially generate predictive labels [22]. Despite the

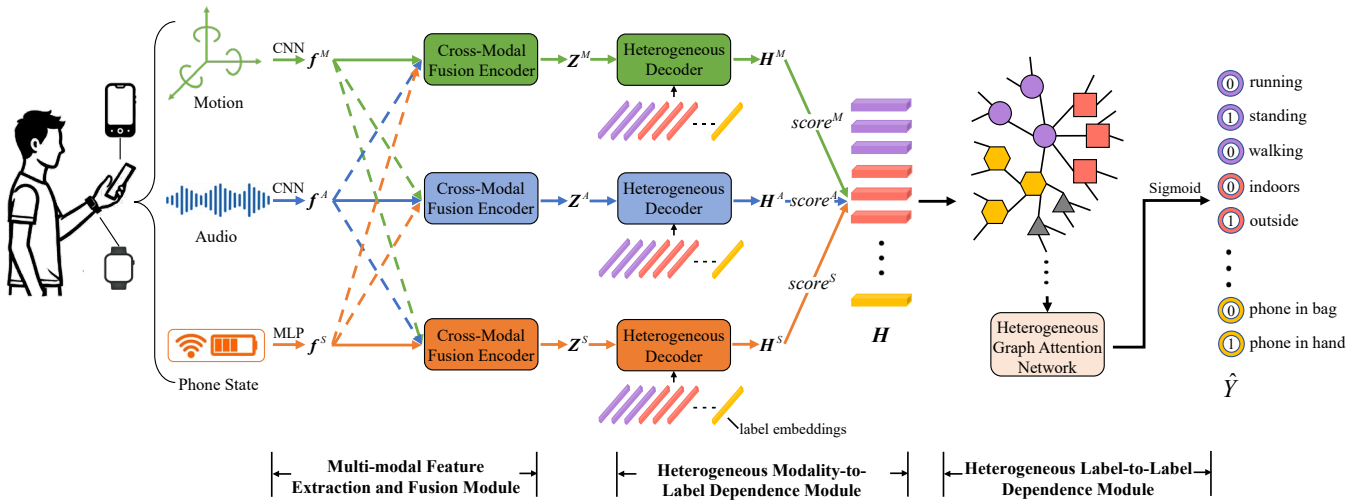


Fig.1 The overall architecture of our proposed HMMBCR network, which takes multi-modal raw sensor data as input, including motion, audio, and phone state data, and has multi-label classification results as outputs, such as simple activity labels ("standing"), place labels ("outside"), and phone position labels ("phone in hand").

significant progress made by these methods, sequentially modelling label correlations may result in error propagation and remain inadequate for modeling complex label dependencies.

In contrast to the aforementioned methods, graph-based approaches have been proven effective in modeling label dependencies. Chen et al. proposed a method that constructed label correlations using a gated graph neural network [23]. Ye et al. applied a dynamic graph convolutional network to learn label-aware representations [24]. Inspired by the success of multi-label graph learning, Mohamed et al. utilized a graph convolutional model to capture the intricate relationships among behavioral context labels [25]. Additionally, Ge et al. proposed a heterogeneous hyper graph network to exploit graphical patterns and internal relationships within labels [26]. Despite the substantial advancements achieved by these studies, they still omit the potential correspondence between modalities and labels.

Recent studies have focused on simultaneously modeling modality-to-label and label-to-label dependencies within a unified framework. Zhang et al. introduced a transformer-based sequence-to-set approach to capture dependencies between different emotion labels and modalities [14]. They further proposed a novel graph message passing network that considers both label-to-label and modality-to-label dependencies in emotion recognition [15]. However, unlike emotion labels, which generally belong to a shared semantic space, behavioral context labels can be categorized by distinct aspects, each with its own semantic structure, giving rise to label heterogeneity. Neglecting this heterogeneity constrains the capacity of existing methods to effectively capture the distinct intra-category and inter-category relationships, leading to reduced accuracy of model prediction. Therefore, addressing label heterogeneity remains a critical challenge in behavioral context recognition.

III. METHOD

In this section, we introduce our heterogeneous multi-modal multi-label behavioral context recognition (**HMMBCR**) network and the overall architecture is shown in Fig 1, consisting of the Multi-modal Feature Extraction and Fusion Module, the Heterogeneous Modality-to-Label Dependence Module and the Heterogeneous Label-to-Label Dependence Module.

A. Task Definition

We first define some notations and formalize the multi-modal multi-label behavioral context recognition task. Let $\mathcal{D}=\{X_i, Y_i\}_{i=1}^{N}$ denote the training dataset, where N is the number of samples in the training set, each sample $X_i=(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, ..., \boldsymbol{x}_i^M)$ is a multi-modal observation of M modalities. And the ground-truth label set $Y_i=\{y_{i,1}, y_{i,2}, ..., y_{i,L}\}$, where $y_{i,l} \in \{0,1\}$ indicates the absence (0) and presence (1) of the l^{th} behavioral context label. The goal of our task is to learn a classification model based on training dataset and apply it to recognize behavioral context labels from unseen samples.

B. Multi-modal Feature Extraction and Fusion Module

In the multi-modal Feature Extraction and Fusion (FEFU) module, features are separately extracted from each modality of the raw sensor data $\boldsymbol{x}^m$. For time-series data (e.g., IMU-based sensor data), 1D-Convolutional Neural Networks (1D-CNN) are employed to capture hierarchical temporal patterns in sequential data and enhance resilience to noise. For discrete state data (e.g., Phone State data), where temporal dependencies are absent, an MLP is utilized to extract nonlinear features, considering the global correlations across the sequence. As a result, we have $\boldsymbol{f}^m \in \mathbb{R}^{l_m \times d_m}$ as depicted in Fig.1 to denote the extracted feature sequence from the m^{th} modality, l_m and d_m are used to represent the sequence length and the feature dimension respectively.

Given the varying sampling frequencies and the complex

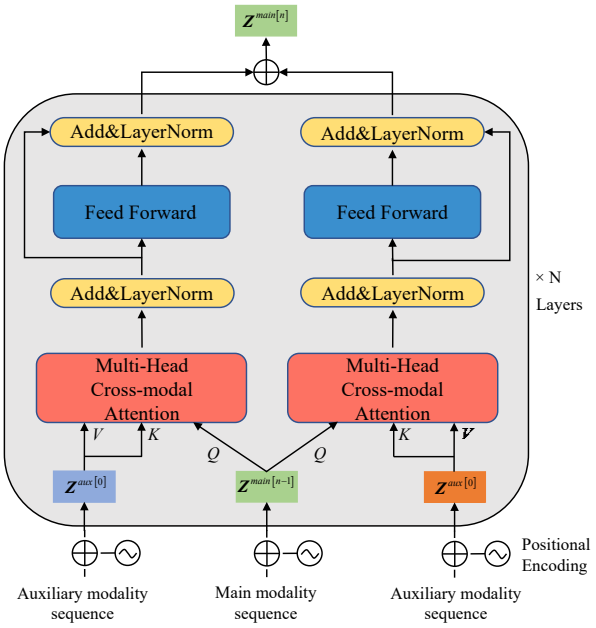


Fig.2 The architecture of cross-modal fusion encoder, illustrating the directional fusion process from auxiliary modality to main modality.

dependencies inherent in features from different modalities, Transformer-based methods are well-suited for capturing inter-modal interactions through the cross-attention mechanism, without requiring strict structural alignment between modalities. Therefore, we leverage a cross-modal Transformer framework [27] to perform sequence fusion between different modalities directionally. Specifically, as illustrated in Fig.2, the cross-modal fusion encoder treats one modality as the main modality, while the remaining modalities serve as the auxiliary modalities. During the fusion process, we first augment the input sequences with positional embeddings to encode temporal order. Then, with each layer of the cross-modal fusion encoder, the low-level features of the auxiliary modalities ($\boldsymbol{Z}^{aux[0]}$) are transformed to a different set of Key and Value pairs, which are used to compute attention weights with Query projected from the intermediate-level features of main modality ($\boldsymbol{Z}^{main[n-1]}$), and the fused modality information is added to get the layer output $\boldsymbol{Z}^{main[n]}$. The cross-attention mechanism allows the model to directly associate relevant elements across unaligned modality sequences. As a result, we have $\boldsymbol{Z}^{m} \in \mathbb{R}^{l_m \times d_m}$ as the encoder output of m^{th} modality after the sequence fusion process.

C. Heterogeneous Modality-to-Label Dependence Module

In this section, we introduce the Heterogeneous Modality-to-Label Dependence (HM2L) module in details. To model modality-to-label dependencies, we first build *M* specialized decoders to generate label representations for each label from *M* modalities. Subsequently, a self-adaptive attention function is employed to dynamically integrate these modalities based on their respective contributions to each label. Furthermore, considering the heterogeneity among different label categories, we adopt a label-categorized strategy to improve contextual relevance and reduce unrelated interference from different modalities to different label categories.

Specifically, after deriving cross-modal fusion encoder output $\boldsymbol{Z}^{m}$, we aim to produce hidden representations for behavioral context labels. Given the diversity of modality data types, such as time-series data and discrete state data, we design specialized heterogeneous decoders to generate label representations from these distinct data types separately.

For $\boldsymbol{Z}^{m}$ belongs to time-series data, we employ the Transformer-based encoder-decoder attention mechanism to obtain label category embeddings $\boldsymbol{H}_c^m \in \mathbb{R}^{r_c \times d_m}$ from $\boldsymbol{Z}^{m}$, as illustrated in Fig.3 (a), where r_c represents the number of labels within category *c*. This design enables the category embeddings to dynamically query relevant information from the encoder output and generate label-specific hidden representations. Unlike grammatical structures in natural language sequences, the label embedding sequences in our task lack inherent positional dependencies. As a result, we omit position embedding and self-attention mechanism in our designed encoder-decoder structure, making it more suitable for capturing unordered label embeddings. To attain the relevant information for each category representations, we denote $Q_c^m = \boldsymbol{H}_c^m W_Q$, $K^m = \boldsymbol{Z}^m W_K$, and $V^m = \boldsymbol{Z}^m W_V$ as the attention operators respectively, where W_Q, W_K, and W_V are the trainable weights. Then the output of encoder-decoder attention sub-layer can be formulated as follows:

$$\hat{\boldsymbol{H}}_c^m = ATT_c\left(Q_c^m, K^m, V^m\right) = \text{Softmax}\left(\frac{Q_c^m \left(K^m\right)^{\mathrm{T}}}{\sqrt{d_m}}\right) V^m, \quad (1)$$

where $ATT_c(\cdot)$ represents the encoder-decoder attention applied to labels within each label category, utilizing a distinct set of projection weights. $\hat{\boldsymbol{H}}_c^m$ represents the output of the encoder-decoder attention block within c^{th} label category in m^{th} modality. Specifically, the operation of heterogeneous decoder for time-series data can be formulated as follows:

$$\begin{aligned} \boldsymbol{H}_c^{m[0]} &= \boldsymbol{E}_c, \\ \tilde{\boldsymbol{H}}_c^{m[n]} &= LN\left(\hat{\boldsymbol{H}}_c^{m[n]} + \boldsymbol{H}_c^{m[n-1]}\right), \\ \boldsymbol{H}_c^{m[n]} &= LN\left(FFN\left(\tilde{\boldsymbol{H}}_c^{m[n]}\right) + \tilde{\boldsymbol{H}}_c^{m[n]}\right), \end{aligned} \quad (2)$$

where $\boldsymbol{H}_c^{m[n]}$ denotes the output of the current heterogeneous decoder, and *n* represents the n^{th} decoder layer. $\boldsymbol{E}_c$ represents the initial embeddings for the c^{th} category. LN and FFN indicates the layer normalization operation and feed forward network separately.

For $\boldsymbol{Z}^{m}$ which belongs to discrete state data that lacks temporal information, we utilize Efficient Channel Attention (ECA) Mechanism [28] to generate category-specific representation as shown in Fig.3 (b). Formally,

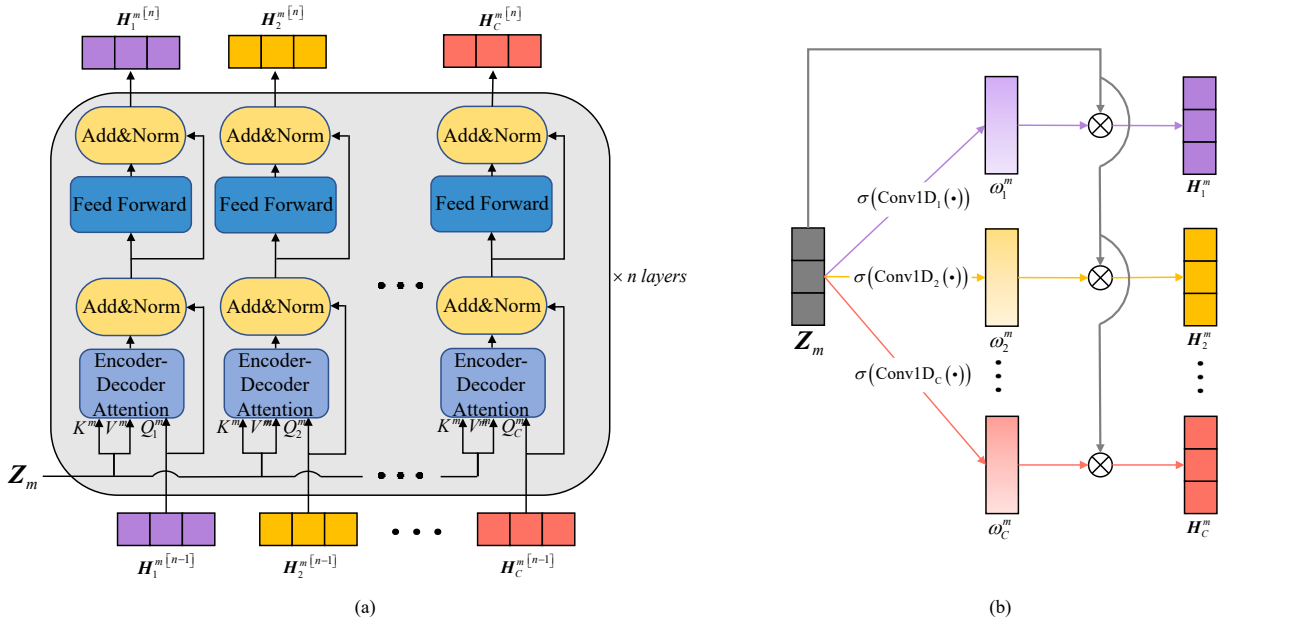


Fig.3 The architecture of heterogeneous decoder for (a) time-series data, and (b) discrete state data

$$\omega_c^m = \sigma(\text{Conv1D}_c(\boldsymbol{Z}^m)),$$
$$\boldsymbol{H}_c^m = \omega_c^m \bullet \boldsymbol{Z}^m, \tag{3}$$

where Conv1D_c denotes the 1D convolution operation for c^{th} label category, ω_c^m represents the category-specific channel attention weights from the m^{th} modality, and σ denotes the activation function, such as Sigmoid. Different from using a single set of projections for all label embeddings, the proposed heterogeneous decoder assigns distinct projection weights to label embeddings of each label category. As a result, labels from different categories could maintain their distinctive representation spaces across all modalities.

Since different modalities contribute differently to a potential behavior context label, after obtaining label-specific generated sequence $\boldsymbol{H}_c^m$, we employ a self-adaptive attention function to dynamically fuse these generated sequences from all modalities using the computed modality-to-label attention matrix. Furthermore, recognizing that labels within the same category may tend to integrate representations from modalities in a similar manner, we assign a distinct set of learning weights to each label category. Formally,

$$score_c = \text{Softmax}\left(\left[\boldsymbol{H}_c^1, \ldots, \boldsymbol{H}_c^M\right] \cdot W_c\right),$$
$$\boldsymbol{H}_c = score_c \cdot \left[\boldsymbol{H}_c^1, \ldots, \boldsymbol{H}_c^M\right], \tag{4}$$

where $W_c \in \mathbb{R}^{d_m}$ represents the trainable weights of category c. $score_c \in \mathbb{R}^{M \times r_c}$ dynamically controls the contribution from different modalities for the c^{th} label category.

D. Heterogeneous Label-to-Label Dependence Module

In this section, we introduce the Heterogeneous Label-to-Label Dependence (HL2L) module in details. To model label-to-label dependencies, we employ a heterogeneous graph attention network (HGAT) [41] to capture the interactions across all label nodes and propagate information (Fig.4).

Since different label categories encode distinct semantic structures and different category-level relationships, treating them uniformly may ignore meaningful relations. To address this, we extend the standard layer-wise graph propagation into a category-aware framework. Information is propagated within each category-specific subgraph, and the added outputs form the final representation. $\boldsymbol{S}$. Formally, the propagation rule at the n^{th} layer is defined as:

$$\boldsymbol{S}^{[n]} = \sigma\left(\sum_{c \in C} \tilde{\text{A}}_c \cdot \boldsymbol{S}_c^{[n-1]} \cdot W_c\right), \tag{5}$$

where $\tilde{\text{A}}_c \in \mathbb{R}^{L \times r_c}$ represents the submatrix of the adjacency matrix $\tilde{\text{A}}$, W_c denotes the category-specific transformation matrix, and $\sigma(\cdot)$ represents the activation function, such as LeakyReLU, and we have $\boldsymbol{S}_c^{[0]} = \boldsymbol{H}_c$ initially.

Due to the differences of specific habits across multiple subjects, the dependencies among behavioral context labels are diversified, making traditional GNN-based models less suitable in this context. Therefore, we compute the adjacency matrix based on Graph Attention Network (GAT) [29]. Furthermore, considering the heterogeneity among different label categories, we expand the single-level attention calculation manner to a dual level way, including label-to-category level attention and label-to-label level attention. Considering a specific node *v*, the label-to-category level attention learns the attention weights from different label nodes to different label categories. Specifically, the category embedding h_c can be denoted as $h_c = \sum_{v'} \tilde{\text{A}}_{vv'} h_{v'}$, the summation of the neighboring label node features $h_{v'}$, where $v' \in N_v$ are the neighborhood nodes of node *v*. Then, the calculation of label-to-category level attention coefficients can be formulated as follows:

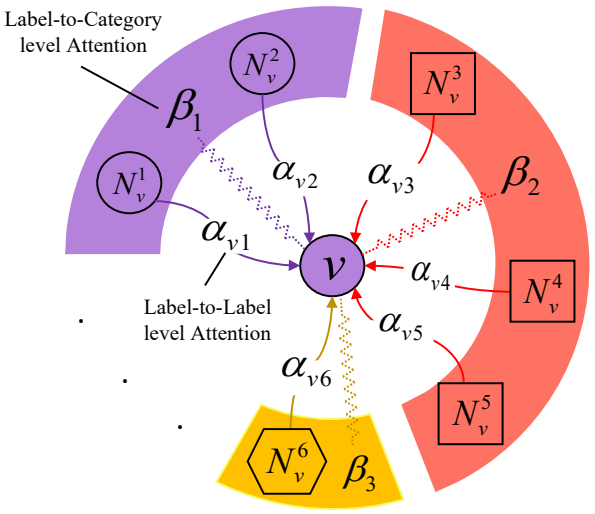


Fig.4 The architecture of heterogeneous graph attention, where the label-to-category attention value β_n is computed between the target node v and each category n, and then the label-to-label attention value α_{vi} is calculated between the target node v and its neighbors, guided by the label-to-category attention β_n.

$$b_c = \sigma\left(\mu_c^{\mathrm{T}} \cdot \left[h_v \,\|\, h_c\right]\right),$$
$$\beta_c = \frac{\exp\left(b_c\right)}{\sum_{c' \in C} \exp\left(b_{c'}\right)}, \tag{6}$$

where μ_c is the attention vector for the category c. To make attention coefficients β_c easily comparable across different categories c, we use Softmax function to normalize them. Once β_c values are obtained, we utilize label-to-label level attention to further determine the significance of different neighboring nodes. Formally, the computation process can be expressed as follows:

$$a_{vv'} = \sigma\left(\lambda^{\mathrm{T}} \cdot \beta_{c'} \left[h_v \,\|\, h_{v'}\right]\right),$$
$$\alpha_{vv'} = \frac{\exp\left(a_{vv'}\right)}{\sum_{i \in N_v} \exp\left(a_{vi}\right)}, \tag{7}$$

where λ is the attention vector, and the v^{th} row v'^{th} column element of A_c is $\alpha_{vv'}$.

E. Behavioral Context Label Prediction

After obtaining $\boldsymbol{S}$, we utilize a prediction function to project each of the L label-specific representations via a projection matrix parameterized by $W^o \in \mathbb{R}^{L \times d_o}$, where each row W_i^o is the learnt output vector for the i^{th} label to predict probabilities $\hat{Y}_i = \left\{\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,L}\right\}$ through a Sigmoid activation function.

Our model is trained based on binary cross-entropy loss function across all behavioral context labels. Given the fact that a certain number of labels are missing due to the uncontrolled data acquisition environment [1], and some rare behavioral context labels are underrepresented in the dataset due to their infrequent occurrence, leading to class imbalance problem, we employ an instance-weighted cross-entropy function to address these issues. Formally,

$$\mathcal{J}_{\text{model}} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{j=1}^{L} \phi_{i,j} \cdot \mathcal{L}_{\text{ce}}\left(\hat{y}_{i,j}, y_{i,j}\right), \tag{8}$$

where N is the number of trained samples, $\mathcal{J}_{\text{model}}$ represents the total objective function used to train the model, $\mathcal{L}_{\text{ce}}$ denotes the cross-entropy loss function, $\phi \in \mathbb{R}^{N \times L}$ is the instance-weighted matrix. When the j^{th} label of the i^{th} sample is missing, $\phi_{i,j}$ will be set to zero to ensure the missing label would not contribute to $\mathcal{J}_{\text{model}}$. Otherwise, $\phi_{i,j}$ will be set according to the proportion of positive label and negative label in each label class.

IV. Experiments

In this section, we provide a comprehensive overview of the experimental settings. To assess the effectiveness of our proposed network thoroughly, we conduct experiments on two public multi-modal multi-label behavioral context datasets: the Extrasensory dataset [10], and the ETRI Lifelog 2020 dataset [30]. To demonstrate the superiority of the proposed model, we compare our model with eight existing competing methods. Furthermore, to validate the effectiveness of each module within our proposed network, ablation studies are performed to illustrate the contribution of each component.

A. Datasets

Extrasensory [10]: This dataset comprises over 300,000 one-minute long recorded instances of behavioral context from 60 users under free-living conditions. The data was acquired by various embedded sensors in smartphones and smartwatches, including an inertial measurement unit (IMU, consisting of accelerometer, gyroscope and magnetometer), a microphone sensor. Additionally, phone state data (such as App running status, battery state, Wi-Fi availability) were also recorded. The IMU records motion modality data at 40Hz. The microphone sensor captures audio modality data in the form of Mel frequency cepstral coefficients (MFCCs) extracted from raw audio signals sampled at 22,050 Hz. The dataset contains 51 refined behavioral context labels such as “standing”, “cooking” and “indoors”, with the possibility of more than one label appearing simultaneously within a single instance. To evaluate the performance of the network on this dataset, we employ a five-fold cross-validation method, where the training and testing folds respectively consist of data from 48 users and 12 users, following the same divisions as in [2].

ETRI Lifelog 2020 [30]: This dataset contains more than 280,000 one-minute recorded behavioral context instances collected from 22 participants during 616 experimental days under free-living conditions. Researchers employed several sensors embedded in smartphones and Empatica E4 wristbands for data acquisition, such as an IMU, a Photoplethysmography

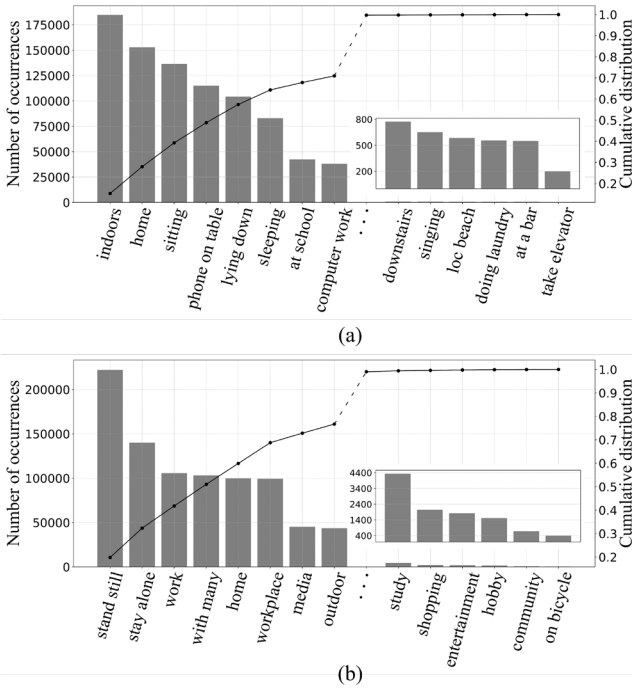


Fig.5 The label counts and distribution of (a) Extrasensory dataset and (b) ETRI Lifelog dataset. We choose the 8 most popular labels and the 6 least popular labels; other labels are represented as suspension points.

(PPG) sensor and a GPS sensor. The IMU embedded in the smartphone records motion modality data at 30 Hz, the PPG sensor embedded in the E4 wristband collects blood volume pressure signal at 64 Hz, and the GPS sensor of the smartphone tracks the longitude and altitude of each participant. These GPS readings are converted into discrete features based on the recorded maximum variation distance and horizontal accuracy within each sample. Since some labels occur extremely infrequently in the dataset, we select 29 behavioral context labels, such as “household”, “home”, “alone”, “walking”. We use a five-fold cross-validation method with a random division.

To illustrate the label imbalance, the label distribution of both benchmark datasets is shown in Fig. 5.

B. Implementation Details

We implement our approach via Pytorch toolkit (torch-1.8.1) in Python version 3.8, with three RTX 3090 Ti GPUs. For both datasets, we train our proposed network in an end-to-end manner for 30 epochs by employing the AdamW optimizer [31], where the batch size is fixed at 512, with learning rate and decay rate set as 0.0001 and 0.02 respectively. Besides, we utilize dropout regularization [32] to avoid overfitting. We use dropout (1) after convolutional block and non-linear mapping function, (2) after getting the attention-probability matrix inside multi-head attention module, and the dropout probability is set to 0.20. By randomly deactivating a certain percentage of activations during training, dropout reduces model capacity and encourages each neuron to learn independently, thereby lowers the risk of overfitting. We also clip the gradients [33] to the maximum norm of 2.0. Further details of the network hyperparameters are available in TABLE 1.

TABLE 1. LIST OF MODEL HYPERPARAMETERS ON EXTRASENSORY DATASET

Stage	Hyperparameters	Values
FEFU	encoder layers	2
	attention heads	4
	dimension d_m	128
HM2L	decoder layers	2
	attention heads	2
	dimension d_m	128
HL2L	graph layers	2
	attention heads	1
	dimension d_m	128
Training	epoch	30
	batch size	512
	optimizer	AdamW
	learning rate	0.0001

C. Evaluation Protocols

Behavioral context datasets are characterized by extreme label imbalance, where many labels are sparse and infrequently recorded [2], thus we recognize that simply using Accuracy as an evaluation metric would be misleading, as it fails to account for infrequently appeared labels. Additionally, this imbalance often leads to class skew that negative instances normally vastly outnumber positive instances within each class. Even a small proportion of False Positives (FP) in the large pool of negative instances can lead to disproportionately low Precision and F1-score, making Precision and F1-score less applicable. Conversely, Recall and Specificity measure the proportion of correctly identified instances within either ground-truth positive or negative class, making them less sensitive to class skew. Finally, we adopt Balanced Accuracy (BA), Recall (REC), Specificity (SPE), and Hamming Loss (HL) as evaluation metrics to fairly and comprehensively evaluate model performance. Formally,

$$BA=\frac{1}{L}\times\sum_{l=1}^{L}\frac{Recall_l+Specificity_l}{2}, \tag{9}$$

$$Recall_l=\frac{TP_l}{TP_l+FN_l}, \tag{10}$$

$$Specificity_l=\frac{TN_l}{TN_l+FP_l}, \tag{11}$$

$$HL=\frac{1}{N*L}\sum_{i=1}^{N}\sum_{l=1}^{L}\mathbb{1}[y_{i,l}\neq\hat{y}_{i,l}] \tag{12}$$

TABLE 2. COMPARISON WITH EXISTING METHODS ON TWO BEHAVIORAL CONTEXT DATASETS

Approaches	Extrasensory				ETRI Lifelog 2020			
	BA (%)	REC (%)	SPE (%)	HL	BA (%)	REC (%)	SPE (%)	HL
BR [34]	67.7±0.9†	66.5±2.9	68.9±1.3	0.33±0.01	50.7±0.3†	48.2±1.9	53.2±1.7	0.49±0.02
CC [35]	68.0±0.9†	66.2±2.7	69.7±1.4	0.32±0.02	50.8±0.4†	48.8±1.9	52.8±1.6	0.51±0.02
Multi-stream [12]	75.0±1.2†	76.7±1.5	73.3±1.6	0.27±0.01	61.7±0.8†	60.3±1.7	63.0±0.4	0.35±0.01
MulT [27]	76.2±0.5†	**79.2±1.9**	73.3±2.1	0.27±0.01	63.4±1.7†	64.2±2.4	62.6±4.2	0.36±0.03
EMOE [46]	76.4±1.1†	76.2±2.4	76.6±0.7	0.24±0.01	63.5±0.8†	63.8±1.8	63.2±1.7	0.35±0.01
SSGRL [23]	76.5±0.8†	77.0±2.2	76.0±1.2	0.25±0.01	63.7±1.3†	63.6±1.6	63.9±3.1	0.34±0.03
Query2Label [36]	75.7±1.3†	75.3±1.6	76.1±2.2	0.25±0.02	63.4±1.2†	63.5±1.7	63.6±2.5	0.34±0.02
SPA-LPR [45]	76.0±1.2†	74.9±1.5	77.2±1.4	0.24±0.01	63.8±1.3†	63.6±1.3	64.0±2.6	0.34±0.02
HHMPN [15]	77.2±0.9†	76.7±1.6	77.6±1.8	0.23±0.01	64.6±0.6†	63.4±2.2	65.9±1.7	0.32±0.01
AMP [37]	77.4±0.7	76.9±2.3	77.9±2.4	0.23±0.03	64.8±1.5†	64.1±1.9	65.5±1.3	0.33±0.01
HMMBCR (Ours)	**78.6±0.4**	78.1±1.1	**79.1±0.9**	**0.21±0.01**	**66.1±0.8**	**65.6±1.7**	**66.6±1.1**	**0.31±0.01**

† indicates the improvement of the HMMBCR over the other methods is significant at the level of p=0.05

where N denotes the number of instances, L is the number of behavioral context labels, TP_l, FN_l, TN_l and FP_l represent true positives, false positives, true negatives and false negatives of the class l, respectively. $1[\cdot]$ is the indicator function returns 1 when the argument is true otherwise 0, $y_{i,l}$ and $\hat{y}_{i,l}$ are the ground truth and predicted value of label l for instance i.

D. Comparison with Existing Methods

To evaluate our approach, we conducted a comparative analysis with 8 existing methods in this section:

1) Binary Relevance (BR) [34]: It transforms the multi-label recognition task into multiple single-label binary classification tasks , which ignores the correlations between labels.

2) Classifier Chain (CC) [35]: It transforms the multi-label task into a chain of binary classification tasks, and takes high-order label correlations into consideration.

3) Multi-stream [12]: It uses a multi-stream CNN to extract features from multi-modal behavioral context data, and then integrates multi-modal features by concatenation, which are subsequently used for the classification .

4) SSGRL [23]: It includes semantic decoupling module and semantic interaction module, where the semantic decoupling module incorporates category semantics to guide learning semantic-specific features, and the semantic interaction module utilizes Gated-GNN to correlate semantic-specific representation.

5) Query2Label [36]: It utilizes label embeddings as queries to check the existence of each label, by performing multi-head cross-attention to pool object features adaptively for the subsequent multi-label classification. This method is considered as the state-of-art in multi-label image classification.

6) SPA-LPR [45]: It proposes a strictly proper asymmetric loss to calibrate multi-label predictions, by promoting confident correct outputs and reducing overconfident errors. Additionally, it regularizes label pair to capture label dependencies. This method is considered as the state-of-the-art approach for multi-label classification.

7) MulT [27]: It utilizes a directional pairwise cross-modal attention mechanism to effectively model the interactions across different modalities, without the need for explicit alignment, which is therefore capable of capturing complex cross-modal relationships.

8) EMOE [46]: It utilizes a mixture of modality experts to dynamically weight different modalities for each multi-modal sample, enabling adaptive multi-modal fusion. Additionally, it employs unimodal distillation to retain the predictive capability of each single modality. This method is considered as the state-of-the-art approach in multi-modal recognition.

9) HHMPN [15]: It utilizes a hierarchical heterogeneous message passing network to address multi-modal multi-label emotion recognition by simultaneously modeling feature-to-label, modality-to-label and label-to-label dependencies. This method is considered as state-of-art in multi-modal multi-label emotion recognition.

10) AMP [37]: It addresses the problems of modality and data biases in multi-modal multi-label emotion recognition. This method employs adversarial temporal masking to balance modality representations by masking dominant emotion-related units, and adversarial parameter perturbation to improve generalization by adding perturbations to model parameters. This method is considered as the state-of-art in this field.

To validate the statistical significance of our approach over other competing methods, we employ one-sided paired Wilcoxon signed-rank test [47], with "†" denoting comparisons in which our approach significantly outperforms the other methods at p=0.05.

E. Ablation Study

To evaluate the effectiveness of the different modules in our approach, we conduct a comprehensive ablation study from 5 variant model compositions.

1) FEFU: This experiment involves using a multi-stream CNN as the automated feature extraction of behavioral context data and a cross-modal interaction module as the fusion of

TABLE 3. RESULTS OF ABLATION EXPERIMENTS ON TWO BEHAVIORAL CONTEXT DATASETS

Approaches	Extrasensory				ETRI Lifelog 2020			
	BA (%)	REC (%)	SPE (%)	HL	BA (%)	REC (%)	SPE (%)	HL
FEFU	76.2±0.5	**79.2±1.9**	73.3±2.1	0.27±0.02	63.4±1.7	64.2±2.4	62.6±4.2	0.36±0.03
FEFU+M2L	76.8±0.8	77.0±2.1	76.6±1.2	0.24±0.01	64.1±1.3	64.6±1.2	63.4±2.9	0.35±0.02
FEFU+M2L+L2L	77.5±0.6	76.9±1.6	78.1±2.2	0.23±0.02	64.8±1.2	64.2±1.7	65.4±2.5	0.33±0.02
FEFU+HM2L+L2L	78.0±0.7	77.4±1.3	78.6±1.4	0.22±0.01	65.4±1.5	65.1±1.8	65.7±1.3	0.33±0.01
FEFU+M2L+HL2L	78.1±0.9	77.8±1.6	78.4±1.8	0.22±0.02	65.6±1.4	65.4±1.6	65.8±2.5	0.32±0.02
Ours	**78.6±0.4**	78.1±1.1	**79.1±0.9**	**0.21±0.01**	**66.1±0.8**	**65.6±1.7**	**66.6±1.1**	**0.31±0.01**

multi-modal data, without the involvement of the HM2L module and the HL2L module. The FEFU component is established based on the existing MulT method [27].

2) FEFU+M2L: On the basis of FEFU, this experiment adds a series of attention-based decoders for the interaction of label-specific embedding and multi-modal data, and a self-adaptive attention function to control the contribution of different modalities. This experiment aims to assess the efficacy of Modality-to-Label Dependence (M2L) module.

3): FEFU+M2L+L2L: On the basis of FEFU+M2L, this experiment adds a GAT subnet to dynamically learn the label correlations. This experiment intends to evaluate the performance of Label-to-Label Dependence (L2L) module.

4): FEFU+HM2L+L2L: On the basis of FEFU+M2L+L2L, this experiment replaces the M2L module with the HM2L module. This experiment is designed to verify the effectiveness of noticing the label heterogeneity in the M2L module.

5): FEFU+M2L+HL2L: On the basis of FEFU+M2L+L2L, this experiment replaces the L2L module with the HL2L module. This experiment is set to examine the capability of noticing the label heterogeneity in the L2L module.

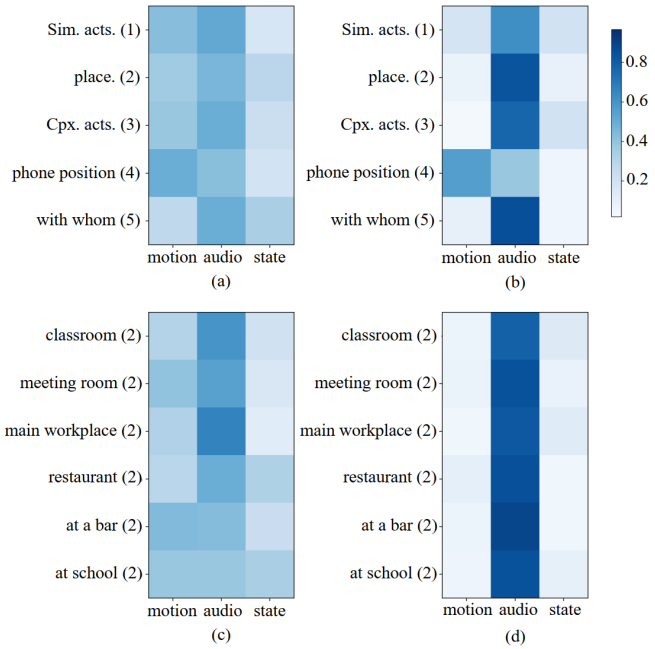


Fig.6 Visualization of modality-to-label dependencies, both **with** and **without** consideration of label heterogeneity, "Sim. acts." is the abbreviation of "simple activities" and "Cpx. acts." is the abbreviation of "complex activities": (a) Dependencies across five label categories **without** consideration of label heterogeneity. (b) Dependencies across five label categories **with** consideration of label heterogeneity. (c) Specific examples of "place" labels to modalities dependencies **without** consideration of label heterogeneity. (d) Specific examples of "place" labels to modalities dependencies **with** consideration of label heterogeneity.

V. RESULTS

A. Comparative Results with Existing Methods

Table 2 shows the performance of various approaches for multi-modal multi-label behavioral context recognition on the Extrasensory dataset and ETRI Lifelog dataset. First, Multi-stream significantly outperforms traditional machine learning methods (BR and CC) on both datasets, indicating that CNN extracts more complex and effective features from behavioral contexts. Second, MulT and EMOE achieves 1.2% and 1.4% higher BA than Multi-stream on the Extrasensory dataset and 1.7% and 1.8% higher BA on the ETRI Lifelog dataset. These results suggest the necessity of modeling interactions among different modalities in multi-modal behavioral context data. Third, the performance of SSGRL, Query2Label and SPA-LPR surpasses that of Multi-stream on both datasets, respectively achieving 0.7%~1.5% higher BA on the Extrasensory dataset and 1.7%~2.1% higher BA on the ETRI dataset. These results indicate that capturing label correlations may bring benefits for the recognition of behavioral context labels. Additionally, HHMPN and AMP achieves 2.2%~2.4% BA on the on the Extrasensory dataset and 2.9%~3.1% BA on the ETRI dataset. These results indicate that simultaneously modelling multi-modal and multi-label attributes is crucial for behavioral context recognition. Finally, our HMMBCR method achieves highest BA among all the competitors on both Extrasensory dataset and ETRI Lifelog dataset, demonstrating the necessity of accounting for label heterogeneity.

B. Results of Ablation Study

The results of ablation study are presented in Table 3, which clearly demonstrate the effectiveness of each module integrated into HMMBCR across two datasets. First, the FEFU+M2L configuration integrates the M2L module into the FEFU module, resulting in a 0.6%~0.7% improvement in BA,

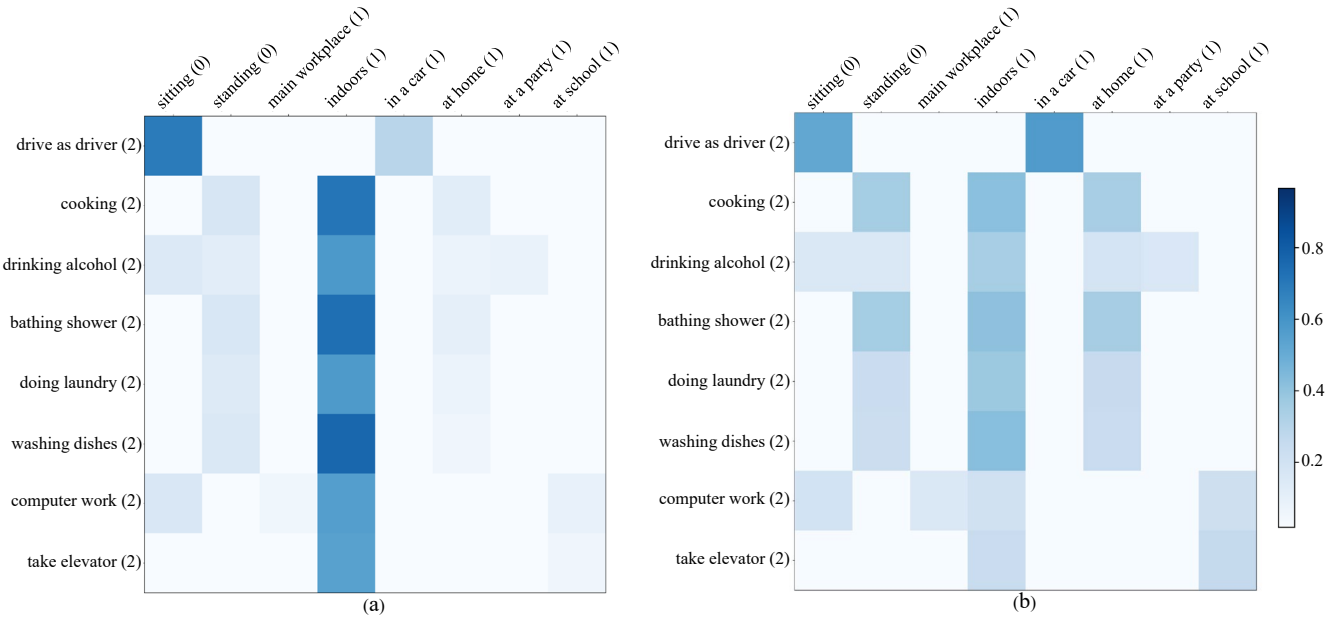


Fig.7 Label-to-Label dependencies visualization **with** and **without** consideration of label heterogeneity: (a) Calculating Label-to-Label dependencies through GAT. (b) Calculating Label-to-Label dependencies through HGAT.

highlighting the utility of effectively capturing modality-to-label dependencies. Second, the FEFU+M2L+L2L module adds the L2L module, further enhancing BA by an additional average increase of 0.7% in BA on both datasets, underscoring the importance of dynamically modeling label correlations. Third, the FEFU+HM2L+L2L configuration replaces the M2L module with the HM2L module, leading to a further increase of 0.5%~0.6% across two datasets, demonstrating the significance of considering heterogeneity among different label categories when modelling modality-to-label dependencies. Finally, the FEFU+M2L+HL2L configuration substitutes the L2L module with the HL2L module, resulting in a 0.6%~0.8% average improvement in BA, emphasizing the importance of considering heterogeneity among label categories when modeling label-to-label dependencies.

C. Visualization

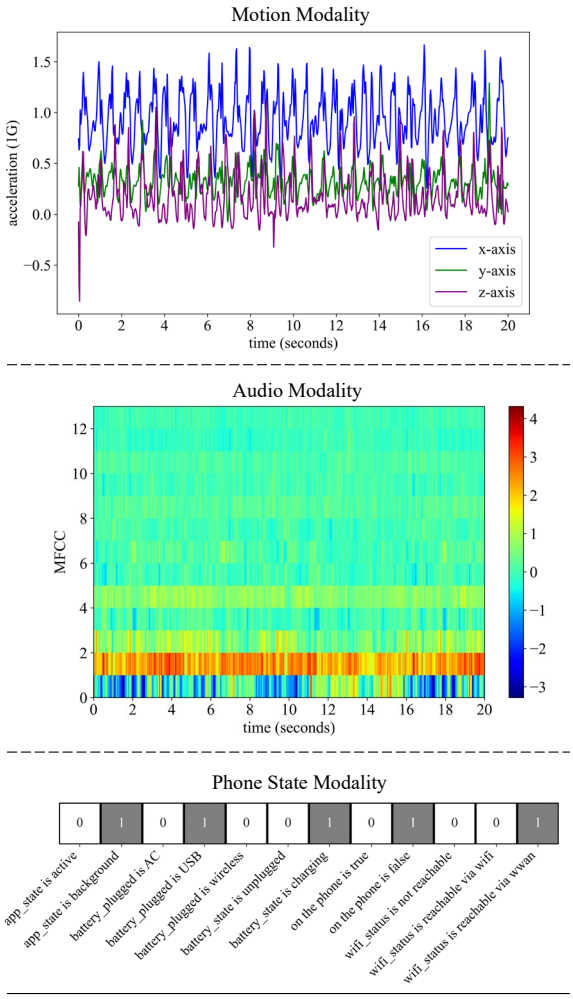
1) Visualization of the Modality-to-Label Dependencies: To illustrate the effectiveness of HM2L, we derive attention weights from both the M2L and the HM2L using the Extrasensory dataset (Fig.6). To provide a comprehensive observation of how each behavioral context category correlate to modalities, we perform average pooling on the modality-to-label attention weights among the labels within the same category, and we use numbers 1-5 to indicate 5 specific label categories (Fig.6 (a) and Fig.6 (b)). Moreover, to understand the relationship between each specific behavioral context label and the corresponding modalities, we derive detailed modality-to-label attention weights for labels within the same category. Due to the space constraint, we select 6 labels within the label category "place" (Fig.6 (c) and Fig.6 (d)).

Through the comparison of Figures 6 (a) and 6 (b), we can observe that accounting for label heterogeneity effectively enhances the attention weights for strongly correlated modalities in behavioral context labels while reducing the attention weights for weakly correlated modalities. For example, from Fig.5 (b), the label category "places" shows a stronger association with the modality "audio", which contains extensive environmental background information.

Additionally, in the M2L (as depicted Fig.6 (c)), these labels exhibit a strong association with the "audio" modality, but this does not indicate a clear predominance to other modalities. While in the HM2L (as depicted Fig.6 (d)), the connection between these labels and the primary modality is strengthened. This observation suggests that considering the heterogeneity in modality-label dependencies helps to achieve consistent attentions for labels within the same category across each modality, while differentiating the attentions of labels from different categories.

2) Visualization of the Label-to-Label Dependencies: To illustrate the effectiveness of HL2L, we compare the differences between Label-to-Label attention weights calculated through the GAT and those calculated through the HGAT on the Extrasensory dataset (Fig.7). Due to space constraints, we select 8 labels from label category "complex activities", 2 labels from label category "simple activities" and 6 labels from label category "places". It demonstrates the directed attention weights calculated from the labels within the category "complex activities" to the labels within the category "simple activities" and the category "place" in the form of attention matrix.

In Fig.7 (a), the attention weights of GAT show that each "complex activity" label is strongly associated with specific "simple activity" or "place" labels. Specifically, the "drive as a driver" label shows a strong connection to the "sitting" label, while other labels within "complex activities" are predominantly linked to the "indoors" label. In contrast, Fig.7 (b) shows that the HGAT-based attention weights are more evenly distributed across potentially relevant nodes. For



True labels: outside, strolling, talking, phone in bag, with friends
MulT: outside, strolling, phone in bag
HMMBCR: outside, strolling, talking, phone in bag, with friends

Fig.8 A case of behavioral context labels predicted by MulT [27] and HMMBCR (Ours) on the Extrasensory dataset.

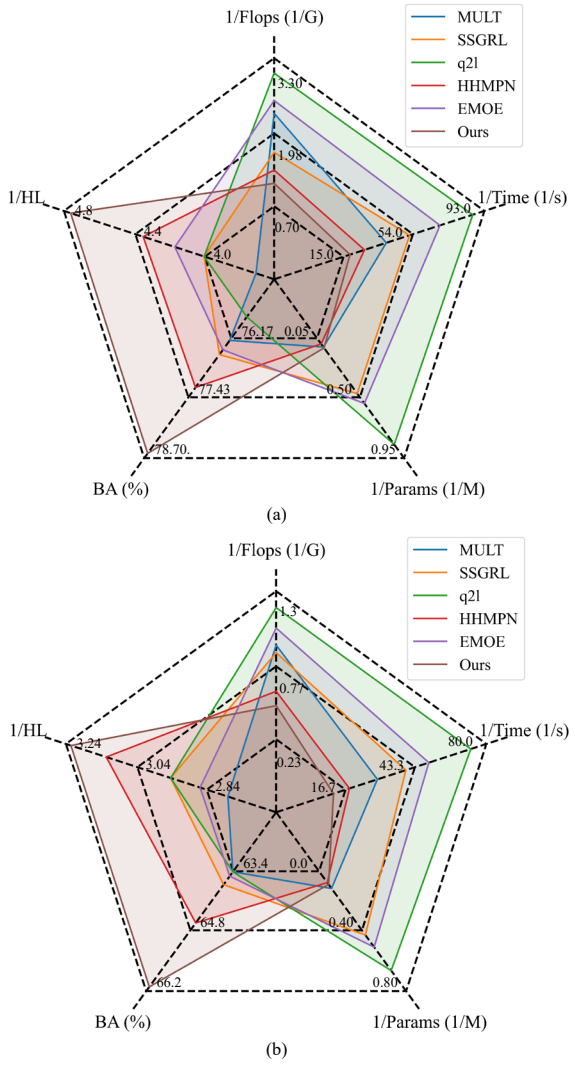


Fig.9 Radar-chart comparison of the proposed HMMBCR with five competing methods in five dimensions: 1/FLOPs, 1/latency time (1/Time), 1/Parameter count (1/Params), Balanced Accuracy (BA), and 1/Hamming Loss (1/HL) on (a) Extrasensory dataset; (b) ETRI Lifelog 2020 dataset.

example, “drive as a driver” is now more evenly connected to both the “sitting” and the “in a car”, and “drinking alcohol” has increased attention to the previously underrepresented “at a party” label. The observations indicate that the heterogeneous architecture adjusts label-to-label dependencies through the use of label-to-category level attention, resulting in more balanced interactions between “complex activity” labels and “simple activity”, as well as “place” labels. This adjustment mitigates the risk of over-reliance on a few strongly connected labels and enhances the comprehensively engagement with relevant labels.

3) Visualization of a behavioral context recognition case: We present a case study using the behavioral context “a person is walking outside, strolling while talking with friends, with his phone in bag”, as shown in Fig.8. While MulT correctly identifies three of the ground truth labels, it fails to recognize “strolling”, “talking” within the “complex activities” category and “with friends” within the “with whom” category. In contrast, our proposed HMMBCR method successfully detects these indistinguishable labels. This comparison result indicates that our approach effectively improves the recognition performance of behavioral contexts by modelling modality-to-label and label-to-label dependencies simultaneously.

D. Analysis of Model Computational Complexity

Figure 9 illustrates the computational complexity of our proposed HMMBCR model in terms of theoretical FLOPs, parameter count (Params) and end-to-end inference latency (Time). Specifically, FLOPs quantify the theoretical floating-point operations required for a single forward pass, parameter count indicates the model’s memory footprint, and latency time denotes the average wall-clock delay. We test these metrics on a piece of RTX 3090 Ti GPU. On the Extrasensory dataset, HMMBCR approach requires 0.92 GFLOPs per sample, incurs 54.9 millisecond latency time, and comprises 8.45 M trainable parameters. On the ETRI Lifelog 2020 dataset, our approach requires 2.12 GFLOPs per sample, incurs 68.7 millisecond latency time, and comprises 11.88 M trainable parameters. And

we have conducted on-device deployment experiments on a Raspberry Pi 5 platform to evaluate inference latency and power consumption. Specifically, our approach achieves an average inference latency of 227 milliseconds with a power consumption of 6.086 Watt on the Extrasensory dataset, and 550 milliseconds with 6.129 Watt on the ETRI dataset.

Although the computational complexity demands of our model exceed those ultra-lightweight architectures designed for on device, real time inference, our primary goal is to maximize behavioral recognition accuracy for downstream cloud-assisted applications like mental health assessment [48], [49]. In such scenarios, sensor data is uploaded in batches (e.g., hourly or daily) to GPU-equipped cloud servers. This cloud deployment decouples model complexity from device constraints, allowing the use of higher-capacity models without compromising user experience. Therefore, the observed model complexity remains acceptable under our deployment assumptions.

VI. Discussion

A. Benefits of Modeling Modality-to-Label and Label-to-Label Dependencies

This study analyses the advantages of simultaneously modeling modality-to-label dependencies and label-to-label dependencies in enhancing behavioral context recognition. The ablation results in Table 3 substantiate this assumption. Moreover, the visualizations in Fig.6 and Fig.7 further validate that the attention weights in both the M2L and the L2L are intuitive and logically coherent.

These observations display that the M2L module effectively assigns different attention weights from each behavioral context labels to each modality, while the L2L module dynamically captures the correlations among all behavioral context labels. These findings indicate that modelling modality-to-label dependencies and label-to-label dependencies provides effective information pathways for the generation and interaction of label-specific representations. This inference aligns with previous studies emphasizing the importance of modeling these dependencies in emotion recognition [14], [15]. However, a distinct characteristic of behavioral context recognition is the inherent heterogeneity of different label categories. This heterogeneity introduces new perspectives for enhancing the performance in this domain, which we will analyze in the subsequent section.

B. Benefits of Heterogeneity Considerations

This study investigates the benefits of incorporating heterogeneity among label categories into the modeling of modality-to-label and label-to-label dependencies, as supported by the ablation results in Table 3. Besides, the visualizations in Fig.6 and Fig.7 provide validation for the effectiveness of the heterogeneous architecture, aligning with our initial hypothesis. These observed performance gains can be attributed to the following factors:

First, when modeling modality-to-label dependencies, the heterogeneous architecture enables explicit constraints on the computation of attention weights for the labels within the same label category. This architecture prevents labels with varying characteristics of different categories from sharing the same projection space, by assigning them into distinct learning subspaces, thus reducing the interference from irrelevant modalities. Similar strategy was also adopted in a previous study [38], which implemented a separated weight projection methodology to manage the propagation of different types of graph nodes in the GNN for modeling heterogeneous structured data. Unlike their method, which focuses on bidirectional node information interaction within the heterogeneous graph, our approach utilizes this strategy for unidirectional information transmission from each modality to the labels. This allows for a deeper exploration of the rich modality-label dependencies arising from label heterogeneity.

Second, when modeling label-to-label dependencies, the heterogeneous architecture introduces more logical constraints and intuitive connections through the dual-level attention mechanism, leading to more balanced interactions between the corresponding nodes within the whole graph. This architecture is inspired by a prior research [39], which demonstrated that the superior constraints achieved by the heterogeneous architecture can capture key information at multiple granularities for short text classification, while mitigating the impact of noisy data. Different from their findings, our observation results indicate that effectively leveraging label heterogeneity can facilitate a richer exchange of information between label nodes, thereby reducing reliance on few strongly connected nodes and leading to more accurate and robust behavioral context recognition.

C. Application on Behavioral Modeling of Mental Health Status Inference

Mental health within modern society has become one of the most pressing concerns [40]. To explore the impact of accurate behavioral context recognition on inferencing mental health status, we utilized mood labels from the Extrasensory dataset [10], annotated hourly and categorized as positive (0) or negative (1) affective states based on the short form of Positive and Negative Affect Schedule (PANAS) [41]. From the dataset's 51 available behavioral context labels, we selected 21 high-frequency labels (such as lying down, walking, at home and exercise) recorded at a minute-level granularity. These minute-level labels were aggregated into hourly frequencies to serve as input features for our analysis. As a result, the processed dataset comprised 275 hourly samples (192 positive and 83 negative), each represented by the aggregated frequencies of the 21 selected behavioral labels.

To better understand how behavioral contexts are related to mood states, we conducted a Mann-Whitney U-test to identify behavioral context labels of significantly different frequencies between positive and negative mood samples. Table 4 highlights labels that showed significant differences, including "home", "exercise", "watching TV", "computer work", and "with friends". These labels intuitively align with distinct emotional states, which is consistent with psychological and behavioral research [42], [43]. For instance, social context like "with friends" were observed more frequently during positive affective states, which aligns with existing literature suggesting that social interactions, particularly with close companions, are strongly associated with positive emotions and well-being.

TABLE 4. MANN-WHITNEY U-TEST FOR SIGNIFICANT DIFFERENCES IN BEHAVIORAL CONTEXT BETWEEN AFFECTIVE STATE

behavioral context	U statistic	p-value (<0.05)
at home	9293.5	0.0138
exercise	7434.0	0.0298
watching tv	8438.0	0.0485
computer work	9117.5	0.0069
with friends	6473.0	0.0025

TABLE 5. RESULT OF PREDICTION FROM RECOGNIZED BEHAVIORAL CONTEXTS TO AFFECTIVE STATE

Approach	Accuracy (%)	F1-score (%)
MulT [27]	63.0	46.3
HHMPN [15]	66.4	51.2
HMMBCR (Ours)	68.9	54.3

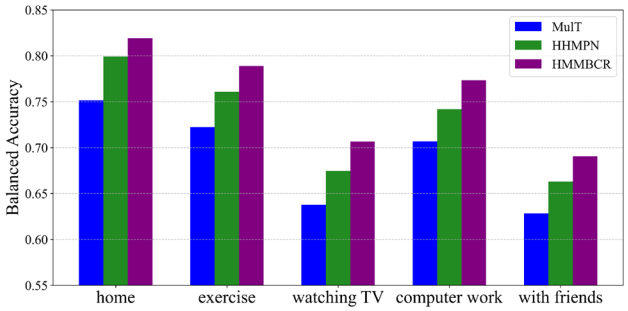


Fig.10 Predicted Balanced Accuracy of behavioral context labels with significant differences between Positive and Negative affective states across three models (MulT, HHMPN, and HMMBCR).

Conversely, cognitive load-related labels like "computer work" were more prevalent in negative affective states, reflecting the well-documented link between work-related tasks and negative emotions such as frustration or anxiety.

To assess the impact of accurate behavioral context recognition on the inference of mental health, we conducted a comparative analysis of our proposed method (HMMBCR) and several competing models in recognizing behavioral contexts. Figure 10 presents the balanced accuracy (BA) across these behavioral context labels of statistically significant differences, demonstrating that our approach consistently outperforms the others on these labels. These labels are particularly difficult to recognize due to their semantic richness and the intricate social and contextual factors involved. For example, recognizing "watching TV" requires integrating information from diverse sensors, such as accelerometers (to detect sedentary behavior), and audio signals (to detect TV-related sounds). In addition, the recognition of "watching TV" is enhanced when contextual labels, such as "at home" and "with friends," are incorporated, because these labels are semantically correlated. By simultaneously modeling modality-to-label and label-to-label dependencies while accounting for label heterogeneity, our approach effectively captures these interdependencies, leading to superior performance in recognizing these intricate contexts.

Building on this improved recognition of behavioral contexts, Table 5 presents the results of mental health inference using a two-layer MLP classifier. The input to the classifier consisted of the hourly frequencies of behavioral contexts recognized by different models, while the output was the binary mood label. Our proposed model achieved the highest accuracy (68.9%) and F1-score (54.3%), outperforming competing approaches such as MulT [27] and HHMPN [15]. These results demonstrate that better recognition of behavioral contexts leads to enhanced mental health status inference performance.

These findings underscore the value of accurate behavioral context recognition in behavioral modeling. By linking fine-grained, minute-level behavioral recognition to hourly affective states, this application provides a reliable framework for understanding the complex relationship between behavioral contexts and affective states. This capability has significant implications for applications in mental health monitoring, personalized interventions, and human-computer interaction, where accurate mental health prediction can drive adaptive and responsive systems

D. Limitations and Future Work

Although the proposed HMMBCR method obtains notable improvements, several limitations that merit future research have been identified.

Firstly, the proposed recognition accuracy oriented multi-modal and multi-label architecture imposes substantial computational and memory demands that might make real-time, on-device inference infeasible on mobile hardware. To broaden applicability, the future work will purse lightweight variants by incorporating efficient modules, such as the Light Graph Transformer, LGT [44] method to reduce computational load and parameter count while preserving robust recognition.

Secondly, many behavioral context labels remain severely underrepresented in benchmark datasets, because exhaustive manual annotation is prohibitively labor intensive, causing annotators to overlook rare but semantically critical labels and leaving sophisticated imbalance-mitigation techniques unable to recover adequate minority class samples. To address this, the future work will implement an active-learning based annotation framework that identifies high uncertainty segments, especially those likely to contain rare behaviors, and dynamically prompts participants to label them. By concentrating human effort where it matters most, this framework aims to collect a more balanced dataset for training robust behavioral context recognition models.

VII. CONCLUSION

In this paper, we proposed a novel heterogeneous multi-modal multi-label approach for behavioral context recognition. Compared with existing behavioral context recognition methods, our proposed approach leverages specialized heterogeneous decoders with self-adaptive attention to model modality-to-label dependencies, and employs HGAT to dynamically model label-to-label dependencies, both accounting for the heterogeneity among label categories. Extensive experimental and visualization results based on two public behavioral context datasets demonstrate the superiority

of our proposed method, which effectively enhances the performance of behavioral context recognition.

REFERENCES

[1] Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches," *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 62–74, Oct. 2017,

[2] Y. Vaizman, N. Weibel, and G. Lanckriet, "Context Recognition In-the-Wild: Unified Model for Multi-Modal Sensors and Multi-Label Classification," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, p. 168:1-168:22, Jan. 2018,

[3] D. Das *et al.*, "Explainable Activity Recognition for Smart Home Systems," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 2, pp. 1–39, Jun. 2023,

[4] A. Khelloufi *et al.*, "A Multimodal Latent-Features-Based Service Recommendation System for the Social Internet of Things," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5388–5403, Aug. 2024,

[5] Q. Meng, B. Liu, H. Zhang, X. Sun, J. Cao, and R. K.-W. Lee, "Temporal-aware and multifaceted social contexts modeling for social recommendation," *Knowledge-Based Systems*, vol. 248, p. 108923, Jul. 2022,

[6] Z. Yu, F. Yi, Q. Lv, and B. Guo, "Identifying On-Site Users for Social Events: Mobility, Content, and Social Relationship," *IEEE Trans. on Mobile Comput.*, vol. 17, no. 9, pp. 2055–2068, Sep. 2018,

[7] R. Wang *et al.*, "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle Washington: ACM, Sep. 2014, pp. 3–14.

[8] K. Mundnich *et al.*, "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Sci Data*, vol. 7, no. 1, p. 354, Oct. 2020,

[9] S. M. Mattingly *et al.*, "The Tesserae Project: Large-Scale, Longitudinal, *In Situ,* Multimodal Sensing of Information Workers," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–8.

[10] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–12.

[11] M. Ehatisham-ul-Haq and M. A. Azam, "Opportunistic sensing for inferring in-the-wild human contexts based on activity pattern recognition using smart computing," *Future Generation Computer Systems*, vol. 106, pp. 374–392, May 2020,

[12] A. Saeed, T. Ozcelebi, S. Trajanovski, and J. J. Lukkien, "End-to-End Multi-Modal Behavioral Context Recognition in a Real-Life Setting," in *2019 22th International Conference on Information Fusion (FUSION)*, Jul. 2019, pp. 1–8.

[13] X. Yang *et al.*, "A Multiscale Cross-Modal Interactive Fusion Network for Human Activity Recognition Using Wearable Sensors and Smartphones," *IEEE Internet Things J.*, pp. 1–1, 2024,

[14] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal Multi-label Emotion Detection with Modality and Label Dependence," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 3584–3593.

[15] D. Zhang *et al.*, "Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing," *AAAI*, vol. 35, no. 16, pp. 14338–14346, May 2021,

[16] J. Ni, H. Tang, S. T. Haque, Y. Yan, and A. H. H. Ngu, "A Survey on Multimodal Wearable Sensor-based Human Action Recognition," Apr. 14, 2024, *arXiv*: arXiv:2404.15349.

[17] S. Bhattacharya, R. Adaimi, and E. Thomaz, "Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–28, Jul. 2022,

[18] H.-T. Vu, M.-T. Nguyen, V.-C. Nguyen, M.-H. Pham, V.-Q. Nguyen, and V.-H. Nguyen, "Label-representative graph convolutional network for multi-label text classification," *Appl Intell*, vol. 53, no. 12, pp. 14759–14774, Jun. 2023,

[19] J. Yuan *et al.*, "Graph Attention Transformer Network for Multi-label Image Classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 4, pp. 1–16, Jul. 2023,

[20] H. Fan, W. Yan, L. Wang, J. Liu, Y. Bin, and J. Xia, "Deep learning-based multi-functional therapeutic peptides prediction with a multi-label focal dice loss function," *Bioinformatics*, vol. 39, no. 6, p. btad334, Jun. 2023,

[21] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[22] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, "Multilabel Image Classification With Regional Latent Semantic Dependencies," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2801–2813, Oct. 2018,

[23] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 522–531.

[24] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, "Attention-Driven Dynamic Graph Convolutional Network for Multi-label Image Recognition," in *Computer Vision – ECCV 2020*, vol. 12366, in Lecture Notes in Computer Science, vol. 12366. , Cham: Springer International Publishing, 2020, pp. 649–665.

[25] A. Mohamed, F. Lejarza, S. Cahail, C. Claudel, and E. Thomaz, "HAR-GCNN: Deep Graph CNNs for Human Activity Recognition From Highly Unlabeled Mobile Sensor Data," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Pisa, Italy: IEEE, Mar. 2022, pp. 335–340.

[26] W. Ge, G. Mou, E. O. Agu, and K. Lee, "Heterogeneous Hyper-Graph Neural Networks for Context-Aware Human Activity Recognition," in *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Atlanta, GA, USA: IEEE, Mar. 2023, pp. 350–354.

[27] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences,"in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2019, pp. 6558–6569.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 11531–11539.

[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio,"Graph attention networks,"in Proc. Int. Conf. Learn. Representations (ICLR), 2018.

[30] S. Chung *et al.*, "Real-world multimodal lifelog dataset for human behavior study," *ETRI Journal*, vol. 44, no. 3, pp. 426–437, Jun. 2022,

[31] I. Loshchilov and F. Hutter,"Decoupled weight decay regularization,"in Proc. Int. Conf. Learn. Representations (ICLR), 2019.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[33] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2013, pp. 1310–1318.

[34] X. Shen, M. Boutell, J. Luo, and C. Brown, "Multilabel machine learning and its application to semantic scene classification," in *Storage and Retrieval Methods and Applications for Multimedia 2004*, SPIE, 2003, pp. 188–199.

[35] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach Learn*, vol. 85, no. 3, pp. 333–359, Dec. 2011,

[36] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2Label: A simple transformer way to multi-label classification,"in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 152–161.

[37] S. Ge, Z. Jiang, Z. Cheng, C. Wang, Y. Yin, and Q. Gu, "Learning Robust Multi-Modal Representation for Multi-Label Emotion Recognition via Adversarial Masking and Perturbation," in *Proceedings of the ACM Web Conference 2023*, Austin TX USA: ACM, Apr. 2023, pp. 1510–1518.

[38] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous Graph Transformer," in *Proceedings of The Web Conference 2020*, Taipei Taiwan: ACM, Apr. 2020, pp. 2704–2710.

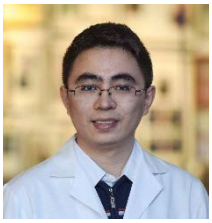
[39] T. Yang, L. Hu, C. Shi, H. Ji, X. Li, and L. Nie, "HGAT: Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification," *ACM Trans. Inf. Syst.*, vol. 39, no. 3, pp. 1–29, Jul. 2021,
[40] D. F. Santomauro *et al.*, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, Nov. 2021,
[41] A. Mackinnon, "A short form of the Positive and Negative Affect Schedule evaluation of factorial validity and invariance across demographic variables in a community sample," *Personality and Individual Differences*.
[42] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, "Context is Everything (in Emotion Research)," *Social and Personality Psychology Compass*, vol. 12, no. 6, p. e12393, 2018,
[43] L. Bechade, G. Dubuisson Duplessis, M. Sehili, and L. Devillers, "Behavioral and Emotional Spoken Cues Related to Mental States in Human-Robot Social Interaction," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, Seattle Washington USA: ACM, Nov. 2015, pp. 347–350.
[44] Y. Wei, W. Liu, F. Liu, X. Wang, L. Nie, and T.-S. Chua, "LightGT: A Light Graph Transformer for Multimedia Recommendation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei Taiwan: ACM, Jul. 2023, pp. 1508–1517.
[45] J. Cheng and N. Vasconcelos, "Towards Calibrated Multi-label Deep Neural Networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27589–27599.
[46] F. Y. Fang, W. K. Huang, G. C. Wan, K. Su, and M. Ye, "EMOE: Modality-Specific Enhanced Dynamic Emotion Experts," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 14314–14324.
[47] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bull., vol. 1, no. 6, pp. 80–83, Dec. 1945.
[48] S. Iranpak, A. Shahbahrami, and H. Shakeri, "Remote patient monitoring and classifying using the Internet of Things platform combined with cloud computing," Journal of Big Data, vol. 8, no. 1, p. 120, 2021
[49] Z. Rashid et al., "Digital Phenotyping of Mental and Physical Conditions: Remote Monitoring of Patients Through RADAR-Base Platform," JMIR Mental Health, vol. 11, p. e51259, Oct.2024

Haodong Liu received the B.Eng. degree in biomedical engineering from the School of biomedical engineering, Sun Yat-sen University, Shenzhen, China, in 2022. He is currently working towards the MS. degree in biomedical engineering with the School of biomedical engineering, Sun Yat-sen University, Shenzhen, China. His research interests include human activity recognition, human behavior context recognition, and multi-modal data application.

Ye Zhang received the Ph.D. degree in Electronic Science and Technology from the National University of Defense Technology (NUDT), Changsha, China, in 2023. He is currently a postdoctoral researcher at the School of Electronics and Communication Engineering, Sun Yat-sen University. His main research interests include human-computer interaction, pattern recognition, time series analysis, and deep learning.

Zhidan Liu received the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2014. After that, he worked as a Research Fellow in Nanyang Technological University, Singapore, and a faculty member with College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is currently an Assistant Professor at Intelligent Transportation Thrust, System Hub, Hong Kong University of Science and Technology (Guangzhou). His research interests include Internet of Things, urban computing, and big data analytic. He is a senior member of CCF, a member of IEEE and ACM.

Dr. Hongyuan Zhu is a senior scientist and PI in the Institute of Infocomm Research (I2R), at the Agency for Science, Technology, and Research (A*STAR), Singapore. He is leading the Advanced Perception Reasoning Lab. He received Ph.D. from Nanyang Technological University (NTU), SG, in 2015. His research interests mainly include multi-modal learning and reasoning. He is the associate editor of Visual Computer since 2020. He also served as the Session Chair of IJCAI 2018 and PCM 2014, Senior Program Committee of IJCAI 2021, and the Guest Editor of IET Image Processing in 2018. He won the Distinguished Paper award of International Consortium of Chinese Mathematicians in 2017 and the Top-10% paper award of ICIP2014. Hs has published around 80 top-tier journal and conference papers, such as CVPR, ICCV, ICML, AAAI, IJCAI, TPAMI, TIP, etc.



Changhong Wang received the B.Eng. and M.Eng. degrees from Harbin Institute of Technology, China, and the Ph.D. degree from The University of New South Wales (UNSW), Sydney, in 2011, 2013, and 2017, respectively. He was a Postdoctoral Research Follow with Baylor College of Medicine, Houston from 2018 to 2020. Now he is an Associate Professor with Sun Yat-sen University, Shenzhen, China. His research interests include wearable and mobile data analysis, and low-power medical electronics design.