2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) | 979-8-3315-2228-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/AINIT65432.2025.11035253

Multimodal information fusion based on event embeddings and spatial-temporal graph convolution networks

Qiuwei Deng
College of Computer Science, Chongqing University
Qingdao Haier Technology Co. Ltd.
National Engineering Research Center of Digital Home Networking
Shandong Key Laboratory of Artificial Intelligence and Nature Interaction in Smart Homes
Chongqing, China
*dengqiuwei@haier.com

Yunlong Tian
Qingdao Haier Technology Co. Ltd.
National Engineering Research Center of Digital Home Networking
Shandong Key Laboratory of Artificial Intelligence and Nature Interaction in Smart Homes
Qingdao, China tianyl@haier.com

Fei Yin
Qingdao Haier Technology Co. Ltd.
National Engineering Research Center of Digital Home Networking
Shandong Key Laboratory of Artificial Intelligence and Nature Interaction in Smart Homes
Qingdao, China
yinfei@haier.com

Wentao Zhang
Qingdao Haier Technology Co. Ltd.
National Engineering Research Center of Digital Home Networking
Shandong Key Laboratory of Artificial Intelligence and Nature Interaction in Smart Homes
Qingdao, China
zhangwentao@haier.com

***Abstract*—The growing diversity of sensor types and information modalities in smart home environments poses significant challenges for effectively fusing heterogeneous data from multiple sources. These challenges stem from variability across sensing channels and substantial differences in data volume, making accurate perception and utilization increasingly difficult. To address this, we proposed UHome Multimodal Fusion Scheme (UHomeMM)—a novel framework that integrates event embeddings with spatial-temporal graph convolutional networks (STGCN) to enhance multimodal fusion in complex environments. UHomeMM fully exploits valuable multi-source heterogeneous information while suppressing irrelevant signals through a cross-channel attention mechanism. Specifically, it encodes sensory data via event embeddings and performs fusion using STGCN to achieve efficient, accurate information integration. Experimental results demonstrate that UHomeMM significantly outperforms traditional methods in fusing heterogeneous data across channels in home environments, effectively reducing noise and enhancing the system's perception accuracy and multimodal feature fusion performance.**

Keywords—multimodal perception, information fusion, spatial-temporal graph convolutional networks, cross-channel attention mechanism, home environment

I. INTRODUCTION

Multimodal fusion technology[1-3] aims to integrate heterogeneous information from different channels into a comprehensive, three-dimensional information base, accurately representing multidimensional data such as "people-devices-environment" in the home. In smart home scenarios, multimodal perception fusion faces several challenges, particularly in multimodal human-computer interaction[4-5], where users interact with the system using speech and image commands. Speech signals are in the form of sound waves, while image signals are presented as pixels, leading to natural morphological differences between these two modalities. Additionally, in multimodal interaction, one channel may dominate the information flow, and the bias of speech or image channels can complicate the fusion process[6].

Traditional multimodal fusion schemes often struggle in complex smart home scenarios. Most existing multimodal techniques are applied to simpler tasks, such as photo albums with textual descriptions or video datasets with labels[7-9], which involve smaller datasets, simpler structures, and limited modality heterogeneity. However, real home environments present more complex and diverse modal information, with significant inter-modal differences.

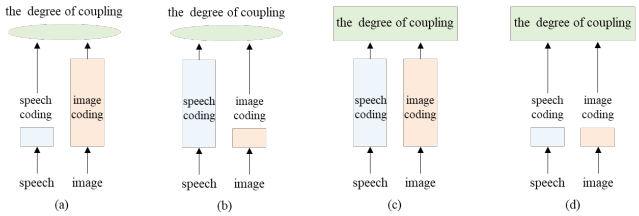

Figure 1 (a) Little speech info, huge image info, weak coupling; (b) Huge speech info, little image information, weak coupling; (c) Huge speech info, huge image info, strong coupling; (d) Little speech info, little image info, strong coupling.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

This complexity and variability have been explored both computationally[10] and experimentally[11]. When only two channels—speech and image—are available, joint multi-channel ideation is illustrated in Figure 1. To provide an intuitive understanding, Figure 1 uses the size of the block diagram elements to represent the information volume of each channel and the degree of coupling between them. As shown, both the complexity and variability of this setup pose significant challenges to the accuracy of multimodal fusion.

To address the challenges of complexity and variability in smart home environments, multimodal perception fusion typically follows two main strategies: early fusion and late fusion. Gunes et al. and Snoek et al. have respectively explored these two approaches for solving cross-channel integration issues[12–13], as illustrated in Figure 2. Early fusion merges features from different modalities at the input stage, and then processes them using a deep learning model. In contrast, late fusion processes each modality independently and combines the outputs using methods such as voting, weighted averaging, or smoothing. After analyzing the strengths and limitations of both methods, Zhang[14] et al. concluded that early fusion is more effective for homogeneous modalities, whereas late fusion is better suited for heterogeneous modalities.

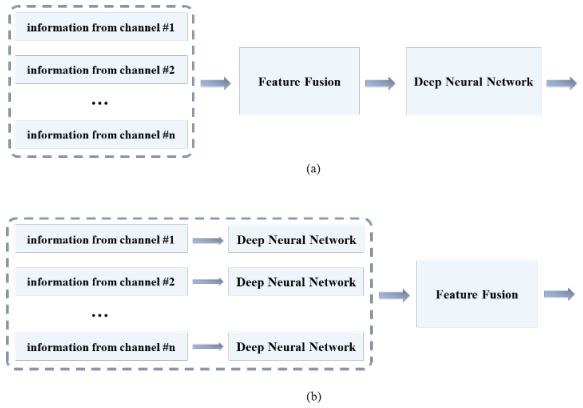

Figure 2. Mechanisms of fusion: (a) Early fusion; (b) Late fusion.

Early fusion techniques are primarily exemplified by the multimodal alignment method introduced by Zhang et al. in 2020, which enhances cross-modal information extraction through a series of alignment strategies. These include attention-based fusion, receptive field fusion via convolution, and kernel-level fusion. Though initially applied to vision-to-vision fusion tasks—such as facial and gesture feature integration—this work marked a meaningful advancement in multimodal fusion research[15].

Kamath et al. further extended early fusion strategies by introducing an end-to-end multimodal interaction system, integrating visual and textual features via attention mechanisms. This was mainly applied to tasks involving image-text datasets, such as photo album collections with accompanying text[16]. Rowan et al. later proposed a federated perceptual fusion approach combining speech, text, and image modalities. Their work demonstrated improved perception accuracy through multi-modal fusion on video datasets annotated with text, further broadening the application of early fusion techniques[17–18]. However, these methods remain largely confined to low-complexity, low-density scenarios and face substantial challenges when applied to dynamic, high-variability environments like smart homes.

In terms of late fusion, a representative method is the Graph Convolutional Network (GCN), which was among the first to apply graph-based structures to complex, heterogeneous data[19]. Building on this, the Graph Attention Network (GAT) introduced by Petar et al. integrated attention mechanisms into graph neural networks, enabling fine-grained fusion across modalities such as speech and vision by capturing associations like morphemes and pixels[20].

Nonetheless, these neural-network-based models often struggle with representing discrete environmental information—such as spatial and temporal context—limiting their effectiveness in more complex environments. To address this, Spatial-Temporal Graph Convolutional Networks (STGCN) were proposed by Han et al. in 2020[21], combining GCN and GAT with temporal-spatial modeling to advance multimodal fusion further. While STGCNs offer significant improvements, they still fall short of fully integrating environmental context within dynamic smart home scenarios.

In summary, smart home environments present unique challenges for multimodal fusion due to the large volume, high heterogeneity, and dynamic variability of sensory data. Misalignment across channels adds further complexity, hampering both convergence and accuracy of fusion algorithms. To overcome these limitations, we propose UHomeMM, a novel multimodal fusion scheme based on event embeddings and an enhanced STGCN. The framework incorporates two core mechanisms: 1) Fusion Coding, which operates at a fine-grained feature level to preserve meaningful representations. 2)

Cross-Channel Attention, which facilitates fusion at a coarse-grained, neural-network level by dynamically adjusting the weight of each channel. Event embeddings are employed to reduce inter-channel heterogeneity, while the STGCN—enhanced by fusion coding and attention mechanisms—enables effective, unified fusion across channels with varying densities and degrees of coupling.

Table 1 provides a comparative summary of recent multimodal fusion approaches and highlights how UHomeMM addresses their limitations, particularly in the context of complex smart home scenarios. Our method significantly enhances fusion accuracy and robustness, offering a scalable and intelligent solution for next-generation smart home systems.

The UHomeMM scheme proposed in this paper is innovative in the following ways:

- The event embeddings used to characterize behavior-based activities in this scheme effectively represent and process action events within the user-behavior relationship graph in the home environment.
- The proposed multimodal fusion method, built upon a STGCN, effectively integrates meaningful heterogeneous information from multiple sources while filtering out irrelevant data. This is accomplished through a multi-channel fusion coding mechanism combined with a cross-channel attention strategy.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

TABLE I. THE MULTI-MODAL MERGING SUMMARY

Method	Brief description	Advantage	Disadvantage
ViLT[11]	Linear mapping and tiling of images and speech, and joint coding of text and picture features	Focusing the main computational effort on multimodal fusion, which in turn improves the speed of model inference	Although the model improves inference speed, it performs mediocrely in terms of accuracy on tasks such as visual quizzing
UNITER[22]	Joint representation learning on graphic data using different feature encoders for image and text	Simple model structure and groundbreaking results on multiple tasks	Performs well on simple tasks but poorly on complex tasks
ALBEF[22]	Multimodal fusion through four pretraining tasks and fine-grained alignment between words and image regions	Multimodal fusion through conditional masking tasks and fine-grained alignment with good performance on some tasks mechanism	The dataset used in this method contains a lot of noise, which may be overfitted to the noise in practical applications, and the generalization ability is average
UHomeMM	Multimodal fusion based on event vector and spatial-temporal graph convolutional networks, event vector to weaken inter-channel heterogeneity, spatial-temporal graph convolution and fusion coding mechanism for fusion of different channels	A unified fusion scheme is built using event embeddings, enhanced spatial-temporal graph convolutional networks, and cross-channel attention to handle information from channels with varying densities and correlations.	Since the method is set in the context of a smart home environment, it also needs to improve the generalization of non-home scenarios

In summary, the proposed scheme effectively integrates multi-source channel fusion coding with the spatial-temporal characteristics of user behavior, while minimizing interference from irrelevant data through an efficient rejection mechanism. Despite the complexity and high heterogeneity of information sources in the home environment, the scheme consistently achieves accurate fusion of meaningful data and demonstrates robust performance in filtering out invalid information.

II. MULTIMODAL FUSION UHOMEMM SCHEME

The UHomeMM scheme proposed in this paper addresses the challenge of fusing heterogeneous information from multiple sources in smart home scenarios. It begins by constructing event embeddings to represent behavior-related events and environment embeddings to capture context-related events, thereby effectively reducing heterogeneity across channels. To further enhance fusion performance, the scheme improves the STGCN by integrating a fusion encoding mechanism and a cross-channel attention mechanism. This enables efficient inter-channel fusion while preserving the unique information of each modality. By considering both the compatibility and semantic integrity of diverse data types, UHomeMM provides a unified solution for fusing multi-source data with varying densities and coupling strengths. The overall technical architecture is illustrated in Figure 3.

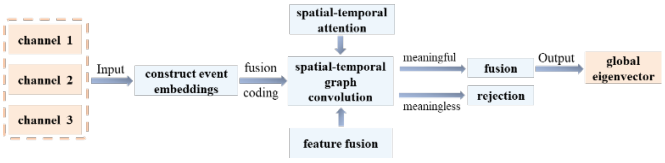

Figure 3. The technical architecture for multi-modal merging scenario.

A. Event Embeddings Construction based on User Behavior Relationship Map

Traditional user behavior modeling methods, such as those based on statistical analysis, typically predict user actions by analyzing historical data patterns. However, these approaches face several notable limitations. First, they struggle to capture the nonlinear dependencies inherent in user behaviors. Second, they often lack accuracy, generalizability, and interpretability, especially when applied to large-scale, sparse datasets.

To better characterize user behavior, this paper constructs a User Behavior Relationship Graph based on in-home behavioral patterns, enabling the vectorized representation of behavioral events. Specifically, each user activity within the household is defined as an event. The definition of atomic events forms the foundation of this approach—we define atomic events as the smallest indivisible units of user actions in a smart home environment. For example, pressing a switch, opening a door, or issuing a specific voice command can each be considered an atomic event. This representation decomposes complex behavioral sequences in the home environment, thereby simplifying the process of vectorizing behavioral events to some extent.

Once atomic events are defined, we construct event sequences based on their temporal order and causal relationships. Suppose a user performs a set of consecutive actions in the home—these actions form a chained event sequence *L (A, B, C, ...)*, as illustrated in Figure 4. The causal relationships between events can be determined using domain knowledge and statistical analysis of large datasets. For instance, turning on a light switch typically precedes the illumination of the room. Such temporal sequences and causal dependencies serve as key foundations for constructing the event sequences.

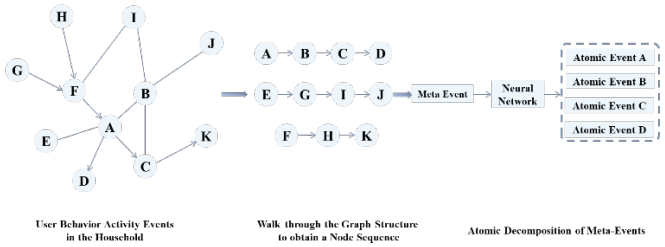

Figure 4. Atomization of sequence of events.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

We begin by representing atomic events as embedding vectors, generated using well-established pre-training mechanisms commonly used in the field of artificial intelligence. These embeddings are then concatenated to form event-to-vector (event2vec) representations, effectively encoding composite behavioral events as continuous vector sequences. Subsequently, these event-level representations are further aggregated into node-to-vector (node2vec) embeddings, forming a graph structure $G(V,E)$ where V corresponds to a behavioral event and E denotes a relationship—typically causal or temporal—between events.

Mathematically, a sequence of user behaviors within the household can be modeled as a pseudo-random walk on the graph G. Unlike a purely stochastic process, this walk is guided by causal dependencies between atomic event embeddings. These constraints ensure that the traversal path respects the natural causality and progression of user activities, thereby preserving the semantic integrity of behavior sequences within the smart home environment. The graph structure constructed in this manner effectively captures the causal dependencies of user behaviors. Moreover, considering that user activities in real-world smart home scenarios exhibit clear temporal and spatial periodicity and correlations, the STGCN offers advantages over other graph neural networks such as GCN. By introducing additional temporal and spatial attention modules, STGCN explicitly computes attention scores for behavioral sequences across the spatiotemporal dimensions. This allows it to better uncover the correlations of user behaviors in both time and space, ensuring that event embeddings accurately reflect the spatiotemporal relevance of user behaviors.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \end{bmatrix} \tag{1}$$

Where, x_i is the embedding vector obtained on each mainstream information medium (communication network, speech, image) representing information of this dimension, and X is the vector matrix obtained by splicing, Figure 5 illustrates the embedding vectors of each node after representation learning is performed on the graph structure.

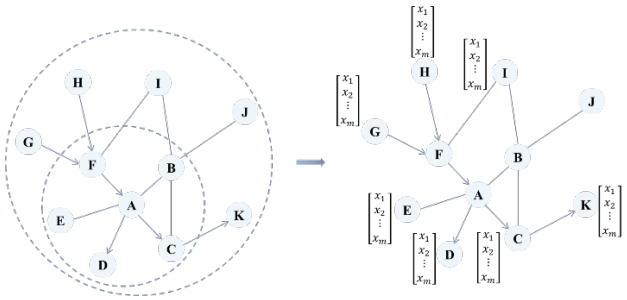

Figure 5. From event2vec to node2vec to node.

If the graph $G(V,E)$ is modeled directly, although it accurately reflects the original information, the nonlinear relationships between the nodes V and the edges E of the graph are not captured. This makes it difficult to apply advanced neural networks and deep learning techniques, which are well-established in the field of artificial intelligence, for further analysis and processing. To address this, the GCN[19], introduced by Thomas et al., was developed, marking the first time that graphs could be analyzed and processed using convolutional neural networks. In this study, we adopt this concept, as shown in Figure 6, where the graph convolution network is implemented by convolving the feature matrix and applying weights to it.

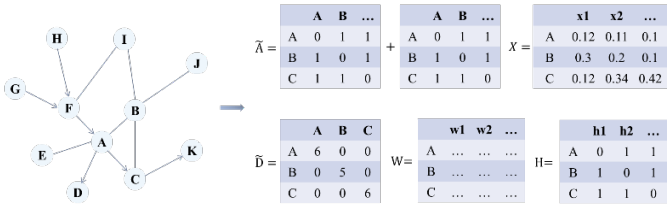

Figure 6. GCN analysis and processing of family behavior activity events.

The basic principles of GCN analysis and processing are:

$$H = \sigma\left\{\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{+\frac{1}{2}}XW\right\} \tag{2}$$

Where, as shown in Figure 6, $\tilde{A}$ is the adjacency matrix, $\tilde{D}$ is the in/out degree matrix, X is the event vector matrix, W is the weight matrix reflecting the relationship of the events with each other, and H is the output feature after processing by graph convolutional network.

The resulting feature matrix H, which represents the user's behavioral activities in the home, is derived from the mainstream information media (communication networks, speech, images), portrays the causal dependencies of behavioral activity events on each other, and is acquired through pre-training mechanisms that are well established in the field of Artificial Intelligence (AI), and is therefore mathematically sound, and can be utilized for further analysis and processing using neural networks and deep learning, which are well established in the field of AI. analysis and processing, for example, it can be used to learn the event labels, or used as input features for federal decision making in Bayesian networks.

B. Multimodal Sensory Information Fusion based on STGCN

The proposed multimodal perception fusion method UHomeMM consists of two main modules: the fusion of multi-source channel features and the fusion of spatial-temporal information. First, the fusion of multi-source channels takes the multi-source features processed by UHome Multimodal Noise Enhancement (UHomeMDE) as input. These features are then fused through a multimodal coding process using the encoder module of a transformer (Transformer Encoder). Second, the fusion of spatial-temporal information is based on the User Behavioral Event Network (UBEN), represented as $G(V,E)$, constructed in the previous section. Temporal and spatial attention weights are calculated, with the spatial non-Euclidean features captured using a GCN. Additionally, a gating mechanism is employed to adaptively fuse the spatial-temporal information.

In the multi-channel fusion module, determining the correlation weights among features from different channels is a critical step. We use Mutual Information (MI) as a metric to

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

measure the correlation between channel features of different modalities. Mutual Information quantifies the degree of dependency between two random variables—in this case, the correlation between features from different modalities such as speech and image channels. By computing the mutual information between these modality-specific features, we obtain the initial feature correlation weights.

At the same time, considering that different scenarios and variations in user behavior can affect feature correlations, we designed a dynamic attention weight adjustment mechanism. Specifically, we introduce a Recurrent Neural Network (RNN)-based weight adjustment module, which takes the user's historical behavior sequences and current contextual information as input. The RNN models the user's behavioral dynamics over time, capturing evolving patterns. Combined with current scene information—such as time, location, and device usage status—the RNN outputs an adjustment factor. This factor is then multiplied with the initial feature correlation weights to produce dynamically adjusted attention weights.

In the spatiotemporal information fusion module, it is also essential to consider the correlation weights of different channel features across temporal and spatial dimensions. Leveraging a spatiotemporal attention mechanism, we calculate attention weights not only over space and time separately but also based on the distribution characteristics of different channel features in both dimensions. For example, we evaluate the importance of speech signals at different time slots and spatial positions, as well as the relevance of image features across regions and time periods. These are computed through specific spatiotemporal attention strategies to generate more fine-grained channel correlation weights.

By combining the above approaches—initial weight estimation using mutual information and dynamic adjustment through the RNN module—the cross-channel attention mechanism can better adapt to varying scenarios and user behaviors, thereby enhancing the accuracy and effectiveness of multimodal information fusion. In this paper, continuous events are used to represent user behavior, such as verbal utterances in the audio dimension and gestures in the visual dimension. Event embeddings are extracted through pre-training. Discrete events, representing environmental information (e.g., spatial localization and time), are also embedded using pre-training techniques. To model the coupling and relationships between these two types of events, an attention mechanism is introduced, enabling fusion via splicing and convolution within the STGCN framework.

The overall structure of the multimodal perception fusion module is shown in Figure 7, while Figure 8 illustrates the specific structure for generating continuous event embeddings from multi-source channel features. These figures highlight the multimodal fusion approach based on the STGCN. The model structure and core principles of the method are described in detail below.

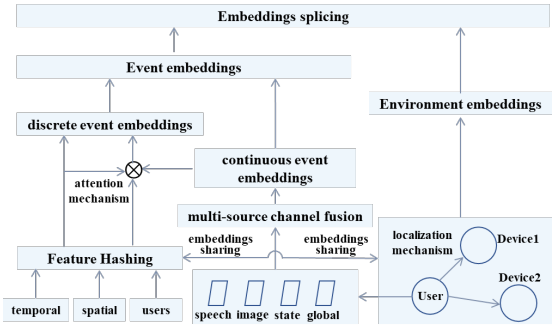

Figure 7. Overview of concatenation and convolution based on STGCN.

First, considering the clear periodicity and correlation of user behavior in real smart home scenarios, both in time and space, STGCN has an advantage over other graph neural networks like GCN. STGCN explicitly computes attention scores for behavioral sequences in both spatial and temporal dimensions by incorporating temporal and spatial attention modules, enabling it to better capture user behavior patterns. Based on these advantages, we choose STGCN as the foundation for introducing an attention mechanism to perform multimodal perceptual fusion of user behavioral sequences across three channels: communication network, speech, and image.

The graph is constructed based on the intra-home user behavior event network $G(V, E)$, described in Section A, where the node set V represents user events, the edge set E represents the relationships between user events, and the feature vector of each node, denoted as X, represents the features from the three channels after noise enhancement and preliminary fusion.

To address the issue of spatiotemporal alignment across different modalities, for speech signals, we first convert them into text and then process the text using common word embedding methods. At the same time, we record the start and end times of the speech, which serve as references in the temporal dimension. For image signals, after applying noise-reduction and enhancement, we use the FCOS object detector to extract regional features. These features are then flattened to form the input sequence for the visual modality, with the timestamp of image capture serving as the temporal reference.

Additionally, this study adopts the encoding strategy from ViLT, where a modality type embedding with values of 0, 1, or 2 is assigned to each modality sequence. This helps the fusion encoder distinguish between different modalities. In this way, data from different modalities are temporally aligned, ensuring consistency in the fusion stage that follows.

In terms of spatial alignment, we use the user behavior event network as the foundation to unify the spatial location information corresponding to each modality. For example, the location of the audio-emitting device is identified for speech signals, while the position of the image capture device is specified for image signals. This approach aligns data from different modalities in the spatial dimension, ensuring spatial consistency and providing an accurate spatiotemporal foundation for the subsequent fusion process.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

The fusion method consists of two main blocks: a multi-channel fusion block and a spatial-temporal information fusion block, as shown in Figure 8 (a) and (b), respectively. For the multi-channel fusion block (a), its function is twofold: first, to initially fuse the noise-enhanced WiFi signal, speech, and image features, mapping them into a common representation space for subsequent integration with spatial-temporal information; second, for the spatial-temporal information fusion block (b), the module leverages the user behavioral network from the previous section. It captures spatial-temporal correlations from home user behavior sequences through the attention-based STGCN. The multi-channel fusion block and the spatial-temporal information fusion block are described in detail below.

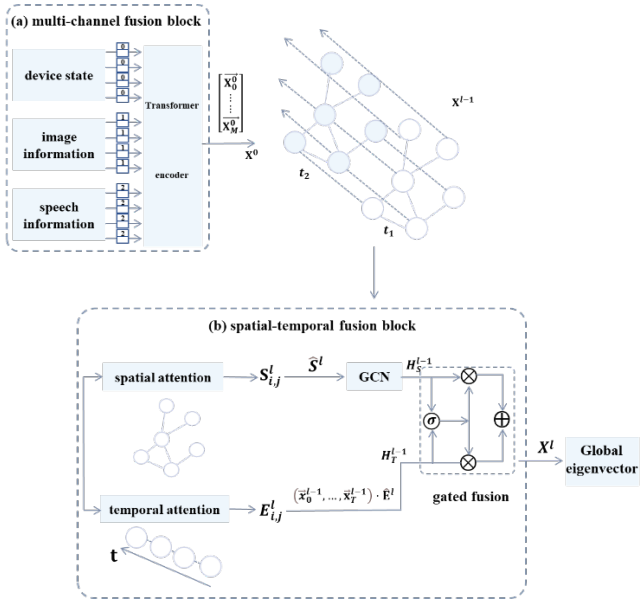

Figure 8. The model architecture of multi-channel fusion block(a) and spatial-temporal fusion block(b).

1) Multi-channel Fusion Block

For the multi-channel fusion block, its primary function is to initially fuse the multi-source channels that have been enhanced through noise reduction, and use the resulting fused embedded representations X as the node feature representations of the user behavior network $G(V, E)$.

The inputs to this module are the multi-source feature representations processed by the UHomeMDE technique, denoted as $x_{State} \in \mathbb{R}^{|V|\times F_1}, x_{Speech} \in \mathbb{R}^{|V|\times F_2}, x_{Vison} \in \mathbb{R}^{|V|\times F_3}$, where $|V|$ represents the number of nodes in the constructed user behavior network, and F_1, F_2, F_3 are the processed feature dimensions of the three channels, respectively. Specifically, the speech signal is converted to text and processed using a common word embedding technique, while the image signal, after noise immunity enhancement, has regional features extracted through the FCOS target detector. These features are then formatted into the input sequence for the visual modality, as shown in Figure 8(a). After the above feature extraction and processing steps, the encoder module of the Transformer is employed to perform modal fusion of the multimodal feature vectors. Finally, the output of the multi-channel fusion module, $X \in \mathbb{R}^{F\times|V|\times T}$ serves as the node input feature representation for the user behavior graph, which is then used for subsequent modeling of spatial-temporal correlations.

2) Spatial-Temporal Information Fusion Block

The spatial-temporal information fusion block is designed to capture non-Euclidean features with spatial-temporal dependencies in real home scenarios, using the constructed user behavioral event graphs. It compensates for the limitations in spatial-temporal data modeling from previous approaches by employing an improved STGCN and a gating fusion mechanism. This allows the module to effectively model and integrate spatial-temporal information, enhancing the overall fusion process.

In terms of model structure, in order to better sense fusion of multi-source channels with spatial-temporal data, so multiple spatial-temporal information fusion blocks need to be stacked in the link, as shown in the module of Figure 8(b). Therefore, mathematically, the input of the lth fusion block is denoted as $X^l = \{\vec{x}_0^l, \dots, \vec{x}_{|V|}^l\} \in \mathbb{R}^{F\times|V|\times T}, l = 0, \dots, L$, where $\vec{x}_i^l \in \mathbb{R}^{F\times T}$ is the F-dimensional feature vector of the ith node in the lth fusion layer that contains the T time step. In particular, the output of the multi-channel fusion module is the initial input feature vector of the node in the user behavior network $G(V, E)$, denoted as X^0. Furthermore, in order to capture the temporal correlation and spatial dependence of user behaviors respectively, we start by calculating the Spatial attention and Temporal attention, respectively, by the following method:

$$\mathbf{S}^l = \mathbf{W}_s \cdot \mathrm{LeakyReLU}\big((X^{l-1}\mathbf{W}_1)\mathbf{W}_2(\mathbf{W}_3 X^{l-1})^T + \mathbf{b}_s\big) \quad (3)$$

$$\hat{\mathbf{S}}^l = \big[\hat{s}_{i,j}^l\big]_{N\times N}, \hat{s}_{i,j}^l = \mathrm{softmax}_j\big(\mathrm{S}_{i,j}\big) = \frac{\exp\big(\mathrm{S}_{i,j}^l\big)}{\sum_{r\in\mathcal{N}(i)}\exp\big(\mathrm{S}_{i,r}^l\big)} \quad (4)$$

The expression of spatial attention is shown in Eq. (3) and (4), and the spatial attention coefficient $\hat{s}_{i,j}^l$ quantifies the degree of spatial relevance of the feature vector of node i about node j on the user behavior network. The vector $X^{l-1} = \{\vec{x}_0^{l-1}, \dots, \vec{x}_{|V|}^{l-1}\} \in \mathbb{R}^{F\times|V|\times T}$ is used to represent the input of the l th spatial-temporal fusion block, while $\mathbf{W}_1 \in \mathbb{R}^T, \mathbf{W}_2 \in \mathbb{R}^{F\times T}, \mathbf{W}_3 \in \mathbb{R}^F, \mathbf{W}_s, \mathbf{b}_s \in \mathbb{R}^{N\times N}$ are all learnable parameter matrices. The computation of this attention score employs a combination of additive and bilinear attention, and uses leakage modified linear cells as a nonlinear activation function to enhance the model's representational capabilities, resulting in the spatial attention matrix $\hat{\mathbf{S}}^l \in \mathbb{R}^{N\times N}$. In addition, considering the existence of oversized graph networks, only the set of first-order neighbors $\mathcal{N}(i)$ of node i is summed up when normalizing the attention coefficients in order to alleviate the computational memory as well as to enhance the scalability of the feature representation.

$$\mathbf{E}^l = \mathbf{V}_e \cdot \mathrm{LeakyReLU}\big(((X^{l-1})^T\mathbf{V}_1)\mathbf{V}_2(\mathbf{V}_3 X^{l-1}) + \mathbf{b}_e\big) \quad (5)$$

$$\hat{\mathbf{E}}^l = \big[\hat{e}_{i,j}^l\big]_{N\times N}, \quad \hat{e}_{i,j}^l = \mathrm{softmax}_j\big(\mathrm{E}_{i,j}\big) = \frac{\exp\big(\mathrm{E}_{i,j}^l\big)}{\sum_{j=1}^T\exp\big(\mathrm{E}_{i,j}^l\big)} \quad (6)$$

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

The expression of temporal attention is shown in Eq. (5) and (6), and the temporal attention coefficient $\hat{e}_{i,j}^{l}$ quantifies the correlation between the time slices i and j in the node feature vectors on the user behavioral network. Its calculation is similar to the spatial attention mechanism. Among them,$\mathbf{V}_1 \in \mathbb{R}^N, \mathbf{V}_2 \in \mathbb{R}^{F\times N}, \mathbf{V}_3 \in \mathbb{R}^F, \mathbf{W}_s, \mathbf{b}_s \in \mathbb{R}^{T\times T}$ are all learnable parameter matrices to obtain the temporal attention matrix $\hat{\mathbf{E}}^l \in \mathbb{R}^{T\times T}$. And then the temporal attention coefficients $\hat{e}_{i,j}^{l}$ between any time slices i,j at the l th spatial-temporal fusion block are obtained by normalizing the temporal dimension.

After computing the spatial attention coefficient matrix $\hat{\mathbf{S}}$ and the temporal attention coefficient matrix $\hat{\mathbf{E}}$, the feature vectors need to be further characterized based on the attention coefficients in the temporal and spatial dimensions, respectively.

First, for the introduction of temporal attention, the normalized temporal attention matrix $\hat{\mathbf{E}}^l$ is multiplied by the output X^{l-1} of the feature vector from the previous layer to dynamically adjust the input sequence. Therefore, the output vector of temporal attention H_T^{l-1} can be represented as follows:

$$H_T^{l-1} = (\vec{\mathrm{x}}_0^{l-1}, \dots, \vec{\mathrm{x}}_T^{l-1}) \cdot \hat{\mathbf{E}}^l \tag{7}$$

Next, for the introduction of spatial attention, in order to fully utilize the topological properties of the user event network, a spatial attention mechanism-integrated spatial-temporal graph convolutional network is employed for modeling. Specifically, since the computational cost of the Laplacian matrix eigenvalue decomposition is high when the graph size is large, we uses Chebyshev polynomial approximation for graph convolution as an improvement. The formula can be expressed as: $g_\theta * x = g_\theta(\mathbf{L})x = \sum_{k=0}^{K-1}\theta_k T_k(\tilde{\mathbf{L}})x$. Furthermore, based on this formula, the paper performs the Hadamard product between the Chebyshev polynomial $T_k(\tilde{\mathbf{L}})$ and the spatial attention coefficient matrix $\hat{\mathbf{S}}$, i.e., $T_k(\tilde{\mathbf{L}})\odot\hat{\mathbf{S}}^l$. This operation dynamically adjusts the correlations between nodes in the home user event network. Thus, the graph convolution formula incorporating spatial attention can be represented as follows:

$$H_S^{l-1} = \mathrm{ReLU}(g_\theta * x) = \mathrm{ReLU}\left(\sum_{k=0}^{K-1}\theta_k\left(T_k(\tilde{\mathbf{L}})\odot\hat{\mathbf{S}}^l\right)x^{l-1}\right) \tag{8}$$

Where g_θ is the convolution kernel, $*$ is the graph convolution operator, the parameter $\theta \in \mathbb{R}^K$ is the polynomial coefficient vector, $\tilde{\mathbf{L}} \in \mathbb{R}^{N\times N}$ is the transformed Laplacian matrix, denoted $\tilde{\mathbf{L}} = 2/\lambda_{max}\,\mathbf{L} - \mathbf{I}_N$, λ_{max} is the largest eigenvalue of the Laplacian matrix$\mathbf{L}$, and$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ is the Chebyshev polynomial.

Finally, after obtaining the spatial and temporal attention outputs H_S^{l-1} and H_T^{l-1}, they are adaptively merged using the following gated fusion mechanism:

In particular, for the computation of the gating z, firstly, the spatial and temporal attention representers are merged (H_S^{l-1}, H_T^{l-1}) and linearly transformed by a learnable parameter matrix W_z and a parameter vector b_z. Then z is obtained by correcting the linear unit (ReLU) as an activation function. Secondly, the gated z and$1-z$ are dot-multiplied and summed by each element for H_S^{l-1} and H_T^{l-1} , respectively, to obtain the output of the l spatial-temporal fusion block. This gated fusion mechanism enables adaptive spatial-temporal fusion of multi-source feature representations at each node, thus enhancing the model's representational capability.

$$z = \mathrm{ReLU}\left((H_S^{l-1}, H_T^{l-1})W_z + b_z\right) \tag{9}$$

$$X^l = z\odot H_S^{l-1} + (1-z)\odot H_T^{l-1} \tag{10}$$

The specific flow of the spatial-temporal information fusion module described above is shown in Algorithm 1. The final matrix X represents the user's event vector within the home. It possesses strong mathematical properties, providing a comprehensive depiction of human-computer interaction and the "human-device-environment" relationship. This is achieved by concatenating the event vector, which captures behavioral activities, with the environment embedding, which encodes environmental context. Consequently, X not only reflects user behavior dynamics but also incorporates environmental factors, making it highly effective for modeling interactions in smart home environments.

Algorithm 1: STGCN-based spatial-temporal information fusion approach

Inputs: user behavior graph $G(V,E)$, multi-channel fusion output X^0, layers of spatial-temporal fusion module L, neighbor nodes of node i $\mathcal{N}(i)$, parameter matrix $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_s, \mathbf{b}_s, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{V}_s, \mathbf{V}_e, \mathbf{b}_e, \mathrm{W}_z, \mathrm{b}_z$.
Output: The feature vector X^L after passing through L spatial-temporal fusion modules.
1: for $l = 1 \dots L$ **do**
2: // Calculation of spatial-temporal attention.
3: $\mathbf{S}^l \leftarrow \mathbf{W}_s \cdot \mathrm{LeakyReLU}((X^{l-1}\mathbf{W}_1)\mathbf{W}_2(\mathbf{W}_3X^{l-1})^T + \mathbf{b}_s)$
4: $\hat{\mathbf{S}}^l = [\hat{s}_{i,j}^l]_{N\times N}$, $\hat{s}_{i,j}^l \leftarrow \exp(\mathrm{S}_{i,j}^l)/\sum_{r\in\mathcal{N}(i)}\exp(\mathrm{S}_{i,r}^l)$
5: $\mathbf{E}^l \leftarrow \mathbf{V}_e \cdot \mathrm{LeakyReLU}(((X^{l-1})^T\mathbf{V}_1)\mathbf{V}_2(\mathbf{V}_3X^{l-1}) + \mathbf{b}_e)$
6: $\hat{\mathbf{E}}^l = [\hat{e}_{i,j}^l]_{N\times N}$, $\hat{e}_{i,j}^l \leftarrow \exp(\mathrm{E}_{i,j}^l)/\sum_{j=1}^{T}\exp(\mathrm{E}_{i,j}^l)$
7: $H_T^{l-1} \leftarrow (\vec{\mathrm{x}}_0^{l-1}, \dots, \vec{\mathrm{x}}_T^{l-1}) \cdot \hat{\mathbf{E}}^l$ // Vector characterization based on temporal attention
8: for$i \in V$ **do:** // Vector characterization based on spatial attention
9: $H_S^{l-1} \leftarrow \mathrm{ReLU}(\sum_{k=0}^{K-1}\theta_k(T_k(\tilde{\mathbf{L}})\odot\hat{\mathbf{S}}^l)\,x^{l-1})$
10: end
11: // Gated fusion mechanism:
12: $z = \mathrm{ReLU}((H_S^{l-1}, H_T^{l-1})W_z + b_z)$
13: $X^l = z\odot H_S^{l-1} + (1-z)\odot H_T^{l-1}$
14: end

The fusion of heterogeneous information through concatenation and convolution is shown in Figure 9.

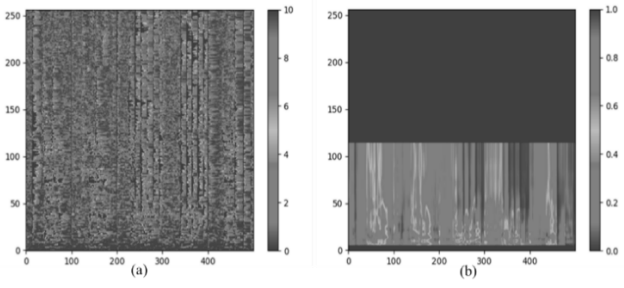

Figure 9. Isomorphic embedding spectrum: (a) speech embedding spectrum;(b) image embedding spectrum.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

The multimodal perception fusion technique homogenizes heterogeneous information, addressing the limitations of single-modal data abundance. This enables different information dimensions to complement and validate each other, resulting in a more comprehensive, multi-dimensional, and accurate understanding of the user's true intent.

C. Rejection of Meaningless Information during Fusion

Accurately recognizing user intent and filtering out interference sources in smart home services is crucial for enhancing user satisfaction. In home environments, voice signals often carry the strongest noise energy and frequent interference, such as sounds from the TV or radio. Therefore, the ability to extract meaningful information from multi-source streams and effectively discard irrelevant or meaningless data is vital for improving the accuracy of intent recognition.

Even with noise reduction, non-noise-related irrelevant information may persist during the fusion process. For example, in voice interactions, while background noise may be suppressed, irrelevant information—such as non-human voices (e.g., TV sounds) or non-interaction commands (e.g., casual conversation or phone calls)—may still remain.

Current rejection methods in the industry typically rely on ASR (Automatic Speech Recognition) for semantic recognition. While this can filter out casual chatter, it struggles to identify non-vocal requests. For example, if a phrase like "I am too hot, turn on the air conditioner for me" is spoken from the TV, the speech recognition system may interpret it as a command to activate the air conditioner based on its semantic text. However, it's clear that this isn't the user's voice, and relying solely on text semantics makes distinguishing non-human voices difficult.

By integrating multimodal information, the UHomeMM scheme enhances the rejection function during perceptual fusion. The system can not only accurately fuse meaningful data but also effectively reject irrelevant information. This improves interaction accuracy, reduces the impact of invalid data, and ensures precise decision-making after fusion. Ultimately, this approach allows the smart home system to interpret user intent more intelligently and efficiently, enhancing the overall service experience.

III. Experimental Design and Analysis

The physical significance of the fused feature vectors can be understood from two perspectives. First, the UHomeMM scheme creates a unified representation by fusing multimodal, multi-source data (e.g., audio and image streams) at the event level, effectively integrating these information streams. Second, it also combines the feature weight vectors, assigning weights to the dimensions of the multimodal data. This process redistributes feature weights within a specific home scenario, based on the relative importance of different segments of the information flow during the current event.

For example, in the kitchen dishwashing scenario, if the user says, "the water is a bit cold," the intent is clearly conveyed. Here, both the image and audio information are highly relevant, so the feature weights for both the voice and image channels are elevated. However, if the system later detects sound from the TV, the relevance of the speech signal decreases, and the weight of the speech channel should be reduced accordingly. This dynamic adjustment reflects the varying importance of different information streams in a given context, providing a theoretical basis for more efficient utilization of multi-source data.

The experimental objectives are as follows: (1) to verify that the feature vector values fused by the UHomeMM algorithm align more closely with the distribution of the weighted original input vectors, and (2) to demonstrate that the fused feature vectors of the UHomeMM algorithm effectively filter out meaningless information. The meaningless information is reflected in the change of the distribution of the weight vector values.

A. Experimental Configuration

1) Data Sets

The dataset used for this experiment includes three publicly available video datasets: VGGSound, GTEA, and EGTEA Gaze++.

- VGGSound consists of 311 categories with over 200,000 videos, totaling 550 hours in duration. The videos in this dataset maintain audio-visual consistency.
- GTEA includes 28 videos, each containing approximately 20 fine-grained action instances, representing seven types of daily activities (e.g., making sandwiches, tea, coffee, etc.), performed by four different individuals.
- EGTEA Gaze++ provides 29 hours of video footage, containing about 15,176 action instances, with a focus on natural kitchen scenes.

After the datasets were summarized, cleaned, and labeled, the data for this experiment was constructed with 150K audio-image pairs, totaling approximately 208 hours, with an average duration of 5 seconds per entry. To increase the dataset size, data augmentation techniques were applied: irrelevant images corresponding to the audio were randomly replaced, the temporal order of images was adjusted, and the temporal sequence of irrelevant audio was shuffled. After data augmentation, the total dataset size increased to 280K entries. Of these, 240K entries were used for the training set, and 40K entries were used for the validation set.

2) Model Setup

In this experiment, the UHomeMM model uses the encoder component of the Transformer architecture as the fusion encoder. The encoder consists of 12 layers of self-attention, with 12 attention heads, a hidden layer dimension of 768, and a total model parameter count of 110 million (110M). For training, the model is run on an Intel Xeon Platinum 8163 CPU (@ 2.50GHz) and a Tesla V100-SXM2-32GB GPU. The AdamW optimizer is used, with an initial learning rate of 0.001 and a warm-up strategy where the warm-up ratio is set to 0.05. The batch size is 64, and the model is trained for 100 epochs.

As illustrated in Figure 10, the input to this experiment consists of two types of information flow features:

- Audio features: The audio is encoded using mono digital-to-analog conversion with a 16kHz frequency sampling rate. The short-time Fourier transform (STFT)

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

has a window length of 32 milliseconds, with a frame shift of 16 milliseconds. The length of the Fast Fourier Transform (FFT) is 512, and the dimensionality of the output features for each frame is 512.

- Image features: Image features are extracted using convolutional operations. The feature dimensions are 256×256. After flattening the image, the resulting feature dimension is 65536.

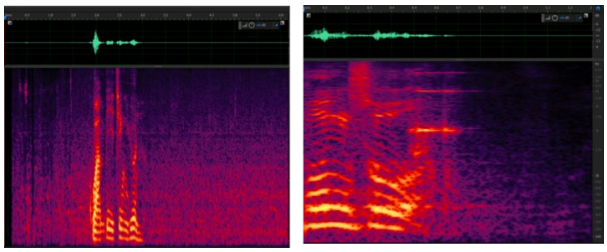

Figure 10. Joint features of texts and images in the constructed data set.

The output of the experiment consists of two vectors: one is the multimodal fusion feature vector, type such as [0.2, 0.4, 0.2, 0.8, 0.75, 0.9, ...]. The second is a weight vector of feature vectors of the form [0.5, 0.5, 0.3, 0.3, 0.3, 0.3, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0, 0, 0]. These vectors represent the fused multimodal features and their respective weights after the information has been processed and integrated by the UHomeMM model.

3) Equipment Environment Settings

A batch of video data was constructed for this experiment, consisting of 80 hours of video footage from kitchen and living room scenes, recorded on the Haier (CED-ECP12N-U5 Black) Smart Home Brain screen in a real smart home environment. The data includes sound and image information related to various activities such as dishwashing, cooking, using the hood and microwave oven, TV background noise, conversations, and phone calls. The spatial layout of the appliances in this scenario is shown in Figure 11.

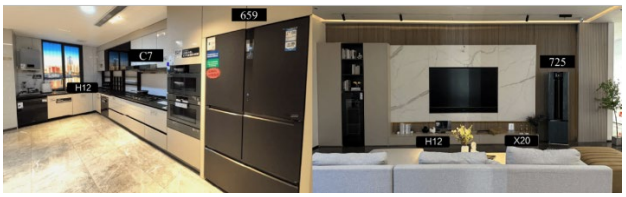

Figure 11. Data sampling environment.

In the living room, the H12 Smart Home Brain screen is equipped with an image capture function, while the X20 speaker and 725 air conditioner are capable of capturing audio. In the kitchen, the H12 Smart Home Brain screen also serves as the image capture device, while the 659 refrigerator and C7 range hood capture audio. This setup enables the collection of multimodal data from both visual and auditory sources, reflecting typical smart home activities and interactions.

a) Experiment Process

The control models used in this experiment are UNITER, Oscar, ViLT, and ALBEF, all of which are widely adopted multimodal pre-trained models in the industry for downstream tasks. However, this experiment does not utilize the full network for downstream tasks; instead, only the first half of each model is employed to calculate the fused feature vectors. The distributions of the fused feature vectors from the different models are then compared to the distributions of the weighted original input feature vectors. The objective function for this comparison is based on the Kullback-Leibler (KL) divergence, which is computed as shown in Eq. (11):

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) log(\frac{p(x_i)}{q(x_i)}) \tag{11}$$

Where $q(x)$ is the approximate distribution and $p(x)$ is the true distribution that $q(x)$ needs to match. This matching process quantifies how much an arbitrary distribution deviates from the true distribution. If the two distributions match perfectly, then the KL divergence will be zero.

$$D_{KL}(p||q) = 0 \tag{12}$$

Otherwise, the KL divergence takes values between 0 and infinity. A smaller KL divergence indicates a better match between the true and predicted distributions. In the fusion experiment, $q(x)$ represents the feature vector calculated by each model, while $p(x)$ denotes the weighted original input feature vector.

The rejection of meaningless information during feature fusion is captured by the generated weight vector. The quality of the weight vector is measured by its distance from the weight label. This distance is typically calculated using the Mean Squared Error (MSE), which is expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \tag{13}$$

where y represents the true label and p is the predicted value.

Since no other model has specifically addressed the task of rejecting meaningless information, no control model is included for this rejection experiment.

b) Analysis of experimental results

The results of the fusion experiments are expressed in terms of KL scatter, as shown in Figure 12.

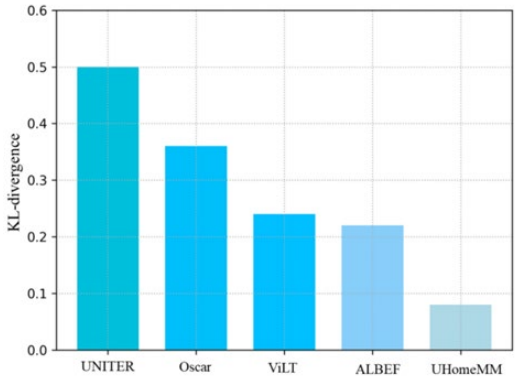

Figure 12. KL divergence comparison results of fusion experiment.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

It can be observed that the UHomeMM algorithm achieves a KL divergence of only 0.08. Since the KL divergence measures the difference between two probability distributions, a smaller value indicates greater similarity between the two distributions. This result suggests that the UHomeMM algorithm effectively aligns the true and predicted distributions, outperforming all other algorithms. The significance of this finding lies in the fact that the proposed algorithm is able to predict the true distribution more accurately.

The loss function of the UHomeMM algorithm during the perceptual fusion training experiment is shown in Eq. (14):

$$loss = \sum_{i=1}^{N} p_i * log(\hat{p}_i) \tag{14}$$

where N denotes the amount of all training data and p denotes the probability of category prediction.

The trend of the loss during the training process is shown in Figure 13. It can be observed that UHomeMM converges faster and achieves a lower final loss value. This is because the UHomeMM algorithm effectively learns from multimodal data, which not only significantly enhances the model's perceptual capabilities but also accelerates learning efficiency, thereby shortening the model's training time.

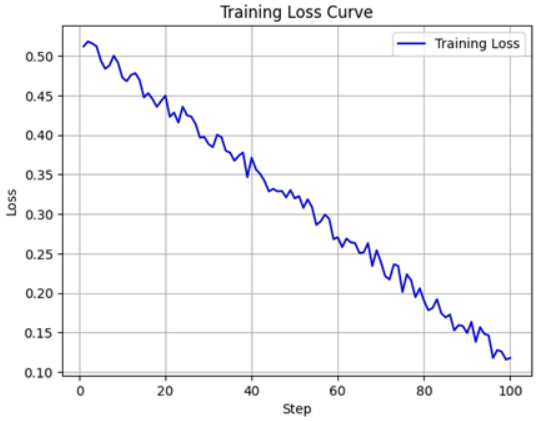

Figure 13. Loss function change curve during the experiment.

The results of the refusal to recognize experiment are expressed as MSE, as shown in Figure 14.

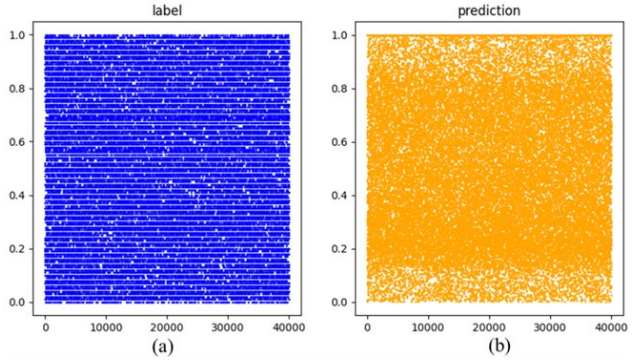

Figure 14. MSE results: (a) Labels for validation set; (b) predictions for validation set.

The average MSE value of the labeling and prediction for the 40K validation set in the rejection experiment is only 0.0234, indicating that the predicted feature weight vector closely matches the labeled feature weight vector. This demonstrates that the experiment has achieved its intended outcome, and the features fused by UHomeMM can effectively filter out meaningless information.

IV. Conclusion

In this paper, we propose the UHomeMM scheme, designed to efficiently fuse perceptual information by integrating multi-source features processed through the upstream UHomeMDE. We introduce a fusion mechanism that combines event embeddings and STGCN, incorporating both fusion coding and cross-channel attention mechanisms. The synergy of these components significantly reduces inter-channel heterogeneity, strengthens the coupling of meaningful information, and effectively suppresses irrelevant data, ensuring the accuracy of multi-source feature fusion. Experimental results demonstrate the effectiveness and robustness of the UHomeMM scheme in handling multimodal information in complex home environments, particularly in enhancing user intent understanding and minimizing the impact of invalid information. Looking ahead, as smart home scenarios continue to evolve, the UHomeMM scheme can be further optimized to address the challenges of more complex and dynamic environments, thus enhancing the intelligence and user experience of smart home systems. In conclusion, the UHomeMM scheme provides a promising solution to the challenge of multimodal information fusion in smart homes, with both significant theoretical and practical implications.

References

[1] Ngiam J, Khosla A, Kim J, et al. Multimodal Deep Learning[C]. The 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[2] Liang P, Zadeh A and Morency L. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions[J]. ACM Computing Surveys, 2024, 56(10): 1-42.

[3] Sahu G and Vechtomova O. Adaptive Fusion Techniques for Multimodal Data[C]. The 16th Conference of the European Chapter of the Association for Computational Linguistics, Kyiv, Ukraine, 2021.

[4] He Y, Cheng R, Balasubramaniam G, et al. Efficient Modality Selection in Multimodal Learning[J]. Journal of Machine Learning Research, 2024, 25(47): 1-39.

[5] Jagnade G, Sable S, Ikar M. Advancing Multimodal Fusion in Human-Computer Interaction: Integrating Eye Tracking, Lips Detection, Speech Recognition, and Voice Synthesis for Intelligent Cursor Control and Auditory Feedback[C]. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023: 1-7.

[6] Gandhi A, Adhvaryu K, Poria S, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information Fusion, 2023, 91: 424-444.

[7] Nagrani A, Yang S, Arnab A, et al. Attention bottlenecks for multimodal fusion[J]. Advances in neural information processing systems, 2021, 34: 14200-14213.

[8] Xue Z, Marculescu R. Dynamic multimodal fusion[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 2575-2584.

[9] Shaikh M, Chai D, Islam S, et al. Multimodal fusion for audio-image and video action recognition[J]. Neural Computing and Applications, 2024, 36(10): 5499-5513.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.

[10] Baltrusaitis T, Ahuja C, Morency L P. Multimodal Machine Learning: A Survey and Taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(2): 423-443.

[11] Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision[C]. International Conference on Machine Learning. PMLR, 2021: 5583-5594.

[12] Gunes H, Piccardi M. Affect recognition from face and body: early fusion vs. late fusion[C]. 2005 IEEE international conference on systems, man and cybernetics: Vol. 4. IEEE, 2005: 3437-3443.

[13] Snoek C G, Worring M, Smeulders A W. Early versus late fusion in semantic video analysis[C]. Proceedings of the 13th annual ACM international conference on Multimedia. 2005: 399-402.

[14] Zhang C, Yang Z, He X, et al. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493.

[15] Zhang C, Jiang M, Zhang X, et al. Multi-Modal Network Representation Learning[C]. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2020: 3557-3558.

[16] Kamath A, Singh M, LeCun Y, et al. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 1760-1770.

[17] Zellers R, Lu J, Lu X, et al. MERLOT RESERVE: Neural Script Knowledge through Vision and Language and Sound[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 16354-16366.

[18] Pan X, Chen P, Gong Y, et al. Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 4491-4503.

[19] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[C]. Proceedings of the 5th International Conference on Learning Representations. Palais des Congrès Neptune, Toulon, France, 2017.

[20] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks[J]. International Conference on Learning Representations, 2018.

[21] Han H, Zhang M, Hou M, et al. STGCN: a spatial-temporal aware graph learning method for POI recommendation[C]. 2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020: 1052-1057.

[22] Chen Y C, Li L, Yu L, et al. UNITER: UNiversal Image-TExt Representation Learning[C]. European Conference on Computer Vision. 2020: 104-120.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 10:15:30 UTC from IEEE Xplore. Restrictions apply.