Adaptive Multimodal Fusion: Dynamic Attention Allocation for Intent Recognition

Bo Hu[1], **Kai Zhang**[2*], **Yanghai Zhang**[2], **Yuyang Ye**[3]

[1]RWTH Aachen University,
[2]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China,
[3]Department of Management Science and Information Systems, Rutgers University
bo.hu2@rwth-aachen.de, kkzhang08@ustc.edu.cn, yhzhang0612@mail.ustc.edu.cn, yuyang.ye@rutgers.edu

Abstract

In recent years, deep multimodal learning has seen significant advancements. However, there remains a lack of multimodal fusion methods capable of dynamically adjusting the weighting of information both within and across modalities based on input samples. In the domain of multimodal intent recognition, the text modality often contains the most relevant information for intent detection, while the audio and visual modalities provide comparatively less critical information. There is a significant variation in the density of important information across different modalities and samples. To address this challenge, we propose a *Dynamic-attention Allocation Fusion* (DAF) method with an adaptive network structure that dynamically allocates attention both within individual modalities and across multiple modalities. This approach enables the model to focus more effectively on the most informative modalities and their respective internal features. Furthermore, we introduce a *Multi-View Contrastive Learning framework based on DAF* (MVCL-DAF). This framework uses distinct and isolated modules to process information from various modalities, taking inspiration from the way the human brain processes multimodal information. Each modality independently infers intent using its respective module, while DAF integrates the multimodal information to produce a comprehensive global intent prediction. The text modality, functioning as the primary modality due to its rich semantic content, guides the other modules in the multi-view contrastive learning process. Extensive experiments demonstrate that our approach significantly outperforms existing state-of-the-art methods.

Code — https://github.com/Freyrlake/MVCL-DAF

Introduction

Intent recognition (Ouyang et al. 2021; Chong et al. 2023) in artificial intelligence involves discerning and identifying the underlying purpose or intention behind a user's input, whether it is conveyed through text, speech, or visual data. Some of the first research in this area mostly focused on one type of input, like text (Zou et al. 2022b; Chong et al. 2023) or visual (Tang et al. 2012; Joo et al. 2014; Aneja et al. 2019). However, in recent years, there have been notable advancements in multimodal intent recognition, which integrates information from multiple modalities to provide a more comprehensive understanding of user intentions.

Figure 1: In intent recognition tasks, the distribution of important information is highly uneven across different modalities and samples.

Multimodal intent recognition extends beyond traditional single-modality approaches by simultaneously analyzing data from text, speech, and visual inputs. This holistic approach allows the model to capture a more enriched and nuanced understanding of the user's intent.

Despite these advancements, existing methodologies frequently fail to account for the varying information density inherent across different modalities and individual samples. As shown in Figure 1, the man in the left image is smiling and greeting the woman with a warm and friendly tone. The pronounced facial expressions and nuanced vocal intonations in this interaction render the acoustic and visual modalities rich with information that can be effectively leveraged to infer intent. Conversely, the image on the right portrays a robot devoid of discernible facial expressions, coupled with a monotonous speech tone, this makes it challenging to extract significant information from the acoustic and visual modalities. Research has demonstrated that textual data often provides more critical insights for intent inference compared to auditory and visual inputs (Han, Chen, and Poria 2021). However, relying exclusively on textual information may not always be sufficient for accurate intent recognition. In certain scenarios, the integration of visual information (e.g., facial expressions and gestures) and acoustic cues (e.g., intonation) is essential for achieving a more precise determination of intent. Nonetheless, discerning intent purely from visual and acoustic data poses challenges due to the often limited

*Corresponding Author.

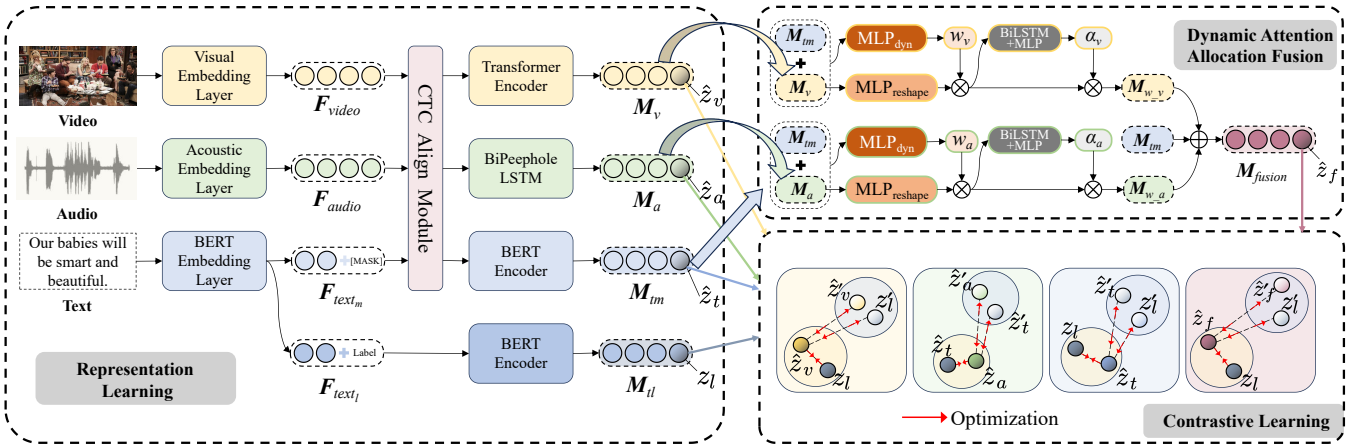

Figure 2: The overview architecture of MVCL-DAF. In DAF, we first compute the acoustic and visual weight matrices based on the input samples by considering the text-acoustic and text-visual pairs. These matrices are used to assign higher weights to the important information within the acoustic and visual modalities. Subsequently, the weighted acoustic and visual high-dimensional vectors are used to calculate the attention scores. These attention scores determine the overall weight of each modality.

and ambiguous nature of these modalities. Therefore, it is imperative to devise strategies that effectively extract pertinent information from auditory and visual stimuli, where intent signals are less evident. Additionally, ensuring that the model's attention is appropriately distributed across the various input modalities is crucial. This approach ensures that the model prioritizes the most informative and relevant sources, thereby enhancing the accuracy of intent recognition.

To tackle the challenge of effective multimodal fusion, we propose a dynamic attention allocation mechanism. This innovative approach amplifies the significance of critical information within the visual and acoustic modalities by adaptively modulating the network architecture during both the training and inference phases. Our method enhances the integration of textual, visual, and auditory inputs by assigning distinct attention weights tailored to individual samples. This allows the model to prioritize the most pertinent information for intent recognition, thereby facilitating more accurate and reliable predictions.

Furthermore, we propose a novel multi-view contrastive learning method for intent recognition (MVCL-DAF), which builds upon the dynamic attention allocation fusion mechanism. This method conceptualizes text, visual, acoustic, and their combined inputs as distinct perspectives for intent recognition. In our framework, high-dimensional representations of textual inputs and intent labels function as anchors in the contrastive learning process, thereby directing the alignment and contrastive learning of other modalities. Inspired by the human brain's processing of visual and auditory information across different regions (Kaas and Hackett 2000; Romo and Salinas 2003; Wandell and Winawer 2015), we employ three independent modules to process each modality separately, enabling the learning of high-dimensional representations tailored to each input type. These representations are then integrated through the dynamic attention allocation fusion module. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art methods across various datasets. The primary contributions of this work are as follows:

- We developed a dynamic attention allocation fusion (DAF) module that adaptively assigns appropriate weights to both the critical information within each modality and the modalities themselves.
- Building upon the DAF module, we introduced a multi-view contrastive learning framework. This method, inspired by the brain's processing of multimodal information, treats different modalities as distinct perspectives. By comparing information across these perspectives, our approach significantly enhances the accuracy of intent recognition.
- We validated our approach on two challenging datasets, demonstrating that it surpasses state-of-the-art methods in the multimodal intent recognition task.

Related Works

Multimodal Fusion Methods

Multimodal fusion refers to the integration of information from multiple distinct modalities or data sources to enhance the accuracy and robustness of computational systems.

Early multimodal fusion methods, such as the Tensor Fusion Network (TFN) (Zadeh et al. 2017), demonstrated how unimodal, bimodal, and trimodal data interact with each other. More recent advancements in multimodal fusion methods have been increasingly based on transformer architectures. For instance, MulT (Multimodal Transformer) (Tsai et al. 2019) is specifically designed to handle unaligned multimodal language sequences by employing directional pairwise cross-modal attention mechanisms. MAG-BERT (Multimodal Adaptation Gate for BERT) extends the conventional BERT architecture to incorporate multimodal data, including visual and acoustic inputs, into the primarily text-based model. DynMM (Xue and Marculescu 2023) conserves computational resources for simpler tasks and more effectively

handles complex data inputs by utilizing a gating function to make real-time decisions at the modality or fusion level. This is supported by a resource-aware loss function that promotes efficiency. Meta's AI research team introduced ImageBind (Girdhar et al. 2023), a model designed for multimodal learning that unifies information across different modalities without the need for explicit alignment or matching. The Deep Equilibrium (DEQ) model (Ni et al. 2024) endeavors to dynamically capture and integrate interactions both between and within modalities by applying nonlinear projections iteratively until an equilibrium state is reached.

Multimodal Intent Recognition

Multimodal intent recognition leverages information from different sources (e.g., text (Zhang et al. 2021b, 2022b, 2019), audio, and visual input) to make more accurate and reliable predictions about user intent. Recent multimodal intent recognition methods such as TCL-MAP (Zhou et al. 2024), consist of two key components: Token-Level Contrastive Learning (TCL) and Modality-Aware Prompting (MAP). The TCL component leverages textual features to guide the learning processes of other modalities, while the MAP module generates prompts that enrich text representations by incorporating insights from video and audio modalities. Additionally, the Contextual Augmented Global Contrast (CAGC) technique (Sun et al. 2024) tackles the challenges of integrating contextual information across multiple videos and aligning modalities to improve intent recognition.

Multi-view Contrastive Learning

In recent years, multi-view contrastive learning has seen significant breakthroughs. The Multi-level Cross-view Contrastive Learning for Knowledge-aware Recommender System (MCCLK) framework (Zou et al. 2022a) leverages both local and global view contrastive learning to enhance recommendation systems that utilize information from knowledge graphs and user-item graphs. By applying multi-level cross-view contrastive learning, this framework effectively captures and exploits the collaborative and semantic relationships within the data, thereby improving recommendation accuracy. FACTORCL (Factorized Contrastive Learning) (Liang et al. 2023), which distinguishes and optimizes both shared and unique information across modalities. This method involves decomposing multimodal information into shared and unique components, which are then optimized through tailored contrastive learning techniques. FACTORCL also incorporates multimodal data augmentations that align with task relevance, enabling the model to learn effectively without relying on explicit labels. This method facilitates the extraction of richer representations that are more pertinent to specific tasks.

Method

Framework Overview

The human brain primarily processes visual information in the occipital lobe (Grill-Spector and Malach 2004), while auditory information is managed by the temporal lobe (Kaas, O'Brien, and Hackett 2012). These two sensory systems interact through multisensory integration processes occurring in several brain regions, including the superior temporal sulcus and the parietal lobe (Stein and Meredith 1993). This integration enhances perception and understanding by combining information from different senses, thereby creating a coherent and unified experience of the environment. Drawing inspiration from these neural mechanisms, we propose a novel Multi-View Contrastive Learning method based on Dynamic Attention Fusion (MVCL-DAF). This method employs three independent modules to process aligned textual, visual, and acoustic data, mirroring the brain's capacity to handle distinct modalities within specialized regions.

Our approach further integrates multimodal information through alignment, dynamic attention allocation fusion, and multi-view contrastive learning modules. This design facilitates a comprehensive, multidimensional understanding of user intent within the model. The MVCL-DAF is composed of two primary components: representation learning and contrastive learning. As illustrated in Figure 2, the representation learning phase encompasses a sequence of steps, including feature extraction, representation learning, modality fusion, and intent recognition. Meanwhile, contrastive learning involves feature extraction, the construction of anchor and positive samples, and the clustering of intent labels.

Multi-modal Representation Learning

Feature Extraction In the intent recognition task, the text modality provides the most crucial information for determining intent. To extract features from the text modality, we use the pre-trained BERT (Devlin et al. 2018) language model as a powerful backbone. For each utterance t and its corresponding intent label l, we concatenate them and employ the BERT tokenizer to derive their token representations: $\mathbf{Z}_{text_l} = [\text{[CLS]}, \mathbf{z}_1, \ldots, \mathbf{z}_{l_t}, \mathbf{l}_1, \ldots, \mathbf{l}_{l_l}, \text{[SEP]}] \in \mathbb{R}^{(l_t+l_l+2)}$. To prevent information leakage during the representation learning process, we create a copy of $\mathbf{Z}_{text_l}$ and mask the label portion, resulting in $\mathbf{Z}_{text_m} = [\text{[CLS]}, \mathbf{z}_1, \ldots, \mathbf{z}_{l_t}, [\mathbf{MASK}]_1, \ldots, [\mathbf{MASK}]_{l_l}, \text{[SEP]}] \in \mathbb{R}^{(l_t+l_l+2)}$. Subsequently, both $\mathbf{Z}_{text_l}$ and $\mathbf{Z}_{text_m}$ are processed through the BERT embedding layer to obtain their respective high-dimensional representations:

$$\mathbf{F}_{text_l} = \text{BERTEmbedding}(\mathbf{Z}_{text_l}), \quad (1)$$

$$\mathbf{F}_{text_m} = \text{BERTEmbedding}(\mathbf{Z}_{text_m}), \quad (2)$$

where $\mathbf{Z}$ denotes the token list. The subscript m indicates that the label has been masked, while l indicates that the label has not been masked. F represents the high-dimensional vector derived through the embedding process. F_{text_l} is used to construct anchor samples in contrastive learning, while F_{text_m} is utilized for representation learning following its alignment with other modalities. F_{text_m} and $\mathrm{F}_{text_l} \in \mathbb{R}^{(l_t+l_l+2)\times d_t}$, where d_t is the embedding size.

The raw visual and acoustic data have been embedded in the MIntRec (Zhang et al. 2022a) and MIntRec 2.0 (Zhang et al. 2024) datasets. The processing methods for the raw video and audio data are described in the Appendix. We directly use the embedded visual and acoustic high-dimensional features provided by the dataset.

$$\mathbf{F}_{video} = \text{Embedding}([f_1, f_2, \ldots, f_{l_v}]), \quad (3)$$

$$\mathbf{F}_{audio} = \text{Embedding}(\mathbf{Z}_{audio}), \quad (4)$$

where f_i denotes the i^{th} frame, l_v is the number of frames. The visual feature embedding, $\mathbf{F}_{video}$, is constructed by concatenating all the individual frame features f_i. Similarly, $\mathbf{F}_{audio}$ corresponds to the audio features, with $\mathbf{Z}_{audio}$ representing the audio sequence. $\mathbf{F}_{audio} \in \mathbb{R}^{l_a \times d_a}$, where l_a is the sequence length and d_a is the audio feature dimension.

Before further processing the information from the three modalities, we first employ the CTC module (Graves et al. 2006) to align the high-dimensional vectors of the three modalities:

$$\{\mathbf{T}_m, \mathbf{V}, \mathbf{A}\} = \text{CTC}(\{\mathbf{F}_{text_m}, \mathbf{F}_{video}, \mathbf{F}_{audio}\}), \quad (5)$$

where $\mathbf{T}_m$, $\mathbf{V}$ and $\mathbf{A} \in \mathbb{R}^{L \times D}$ denotes the standardized features with length L and dimension D. $\mathbf{T}_m$ represents the high-dimensional vector of the text modality after alignment, with the label part masked. $\mathbf{V}$ and $\mathbf{A}$ refer to the aligned high-dimensional vectors for the visual and acoustic modalities, respectively.

We employ three distinct and specialized modules to independently process the textual, visual, and acoustic modalities. The interactions among these modalities are operated within the alignment module, the dynamic attention allocation module, and the contrastive learning module. We use the powerful pre-trained language model BERTEncoder to process textual information with masked labels, and a Transformer encoder (Vaswani et al. 2023) to process visual information. In addressing the acoustic modality, we enhance the traditional Peephole LSTM (Gers and Schmidhuber 2000), enabling it to function comparably to a bidirectional LSTM. This advanced bidirectional Peephole LSTM module is dedicated to the processing of acoustic data:

$$\mathbf{M}_{tm} = \text{BERTEncoder}(\mathbf{T}_m), \quad (6)$$

$$\mathbf{M}_{tl} = \text{BERTEncoder}(\mathbf{T}_l), \quad (7)$$

$$\mathbf{M}_v = \text{TransformerEncoder}(\mathbf{V}), \quad (8)$$

$$\mathbf{M}_a = \text{BiPeepholeLSTM}(\mathbf{A}), \quad (9)$$

where $\mathbf{M}_{tm}$, $\mathbf{M}_{tl}$, $\mathbf{M}_v$, and $\mathbf{M}_a$ represent the encoded text, visual, and acoustic features, respectively. $\mathbf{M}_{tm}$ represents the text modality where the labels have been masked, while $\mathbf{M}_{tl}$ represents the text modality where the labels remain unmasked. $\mathbf{M}_{tm}$, $\mathbf{M}_{tl}$, $\mathbf{M}_v$ and $\mathbf{M}_a \in \mathbb{R}^{L \times D}$. $\mathbf{M}_{tm}$, $\mathbf{M}_v$, and $\mathbf{M}_a$ are ultimately fed into the DAF module for dynamic attention allocation, resulting in the logits scores for classification and the multimodal intent prediction labels for multi-view contrastive learning.

Dynamic Attention Allocation Fusion

Dynamic Attention Allocation Fusion (DAF) possesses an adaptive network structure that dynamically allocates attention both within and across modalities based on the specific characteristics of input samples. In multimodal intent recognition tasks, the text modality typically contains the most crucial information (Zhang et al. 2021a), hence we retain the complete text modality during the fusion process. For the visual and acoustic modalities, dynamic attention allocation is applied. As illustrated in Figure 2, the aligned acoustic and visual feature vectors are concatenated with the text feature vectors, creating new high-dimensional feature representations for text-acoustic and text-visual pairs. These new representations are then fed into a dynamic neural network. The dynamic neural network module has a dynamic depth structure. Specifically, the depth of the network is determined by the characteristics of the input data, rather than being fixed. The dynamic neural network consists of multiple fully connected layers and corresponding gating mechanisms. During initialization, the neural network receives three parameters: input dimension, output dimension, and maximum depth. The network generates multiple linear layers based on the maximum depth. Additionally, the network generates a corresponding gating network for each layer to decide whether to proceed to the next layer. In the forward propagation, after the tensor passes through the linear layer of the current depth, it goes through a ReLU activation function. Subsequently, the gating value for the current layer is computed using a sigmoid function. If the gating value exceeds a predefined threshold, the network recursively calls itself to proceed to the next layer; otherwise, it returns the current output.

The final layer of the dynamic neural network uses the PReLU (He et al. 2015) activation function to enable more flexible expression of the network. The attention weight matrices produced by this module are then element-wise multiplied with the transformed visual and acoustic tensors, assigning greater attention weights to significant information within the visual and acoustic modalities.

$$\mathbf{W}_v = \mathbf{PReLU}(\text{MLP}_{\mathbf{dyn}}(\mathbf{M_v}, \mathbf{M_{tm}})), \quad (10)$$

$$\mathbf{W}_a = \mathbf{PReLU}(\text{MLP}_{\mathbf{dyn}}(\mathbf{M_a}, \mathbf{M_{tm}})), \quad (11)$$

$$\mathbf{F}_v = \text{MLP}_{\mathbf{reshape}}(\mathbf{M_v}), \quad (12)$$

$$\mathbf{F}_a = \text{MLP}_{\mathbf{reshape}}(\mathbf{M_a}), \quad (13)$$

$$\mathbf{F}_{w\_v} = \mathbf{F}_v \circ \mathbf{W}_v, \quad (14)$$

$$\mathbf{F}_{w\_a} = \mathbf{F}_a \circ \mathbf{W}_a, \quad (15)$$

$\mathbf{W}_v$ and $\mathbf{W}_a$ are intra-modal attention matrices for the visual and acoustic modalities, respectively, used to assign greater weights to important information within each modality. $\mathbf{F}_v$ and $\mathbf{F}_a$ are the reshaped high-dimensional vectors for the visual and acoustic modalities. The symbol $\circ$ denotes element-wise multiplication. $\mathbf{F}_{w\_v}$ and $\mathbf{F}_{w\_a}$ represent the element-wise weighted visual and acoustic features.

As shown in Figure 2, $\mathbf{F}_{w\_v}$ and $\mathbf{F}_{w\_a}$ are fed into a BiLSTM and $\text{MLP}_{\text{reshape}}$ to compute the overall modality attention scores for the acoustic and visual modalities. The initially computed values of α_v and α_a are further normalized using L1-normalization, ensuring their sum equals 1.

$$\alpha_v = \mathbf{Sigmoid}(\text{MLP}_{\text{reshape}}(\text{BiLSTM}(\mathbf{F_{w\_v}}))), \quad (16)$$

$$\alpha_a = \mathbf{Sigmoid}(\text{MLP}_{\text{reshape}}(\text{BiLSTM}(\mathbf{F_{w\_a}}))), \quad (17)$$

$$\alpha_a + \alpha_v = 1, \quad (18)$$

$$\mathbf{M}_{w\_v} = \alpha_v \cdot \mathbf{F}_{w\_v}, \quad (19)$$

$$\mathbf{M}_{w\_a} = \alpha_a \cdot \mathbf{F}_{w\_a}, \quad (20)$$

Methods	MIntRec				MIntRec 2.0			
	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
MulT	72.52	71.80	72.60	67.44	56.95	54.26	54.49	40.65
MAG-BERT	72.16	71.30	72.03	67.61	55.87	52.58	53.71	39.93
TCL-MAP	73.69	73.38	73.90	71.59	56.99	54.33	55.07	41.87
MVCL-DAF	**74.72**	**74.61**	**75.07**	**71.94**	**57.80**	**55.05**	**55.82**	**42.03**
Δ	1.03↑	1.23↑	1.17↑	0.35↑	0.81↑	0.72↑	0.75↑	0.16↑

Table 1: Multimodal intent recognition results on the MIntRec and MIntRec 2.0 datasets. We repeated the experiment 10 times with random seeds from 0 to 9. Δ represents the difference between the average results of our method and the highest corresponding metric value among the baselines. The best performance for each metric is highlighted in bold.

The final output of the fusion module is the sum of the dynamically attention-weighted visual and acoustic features and the complete text modality features.

$$\mathbf{M}_{fusion} = \mathbf{M}_{tm} + \mathbf{M}_{w\_v} + \mathbf{M}_{w\_a}. \tag{21}$$

Multi-view Contrastive Learning

In the multi-view contrastive learning component, we use InfoNCE (van den Oord, Li, and Vinyals 2019) as the loss function. The primary idea behind InfoNCE is to maximize the similarity between an anchor and a positive sample (typically an augmented version of the anchor or a related instance) while minimizing the similarity between the anchor and multiple negative samples (unrelated instances).

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\sum_{j=0}^{K} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}, \tag{22}$$

The InfoNCE loss function encourages the model to differentiate between positive and negative samples. In the formula, $\mathbf{z}_i$ represents the feature representation of the anchor sample, while $\mathbf{z}_i^+$ denotes the representation of the positive sample that is semantically related to the anchor. The similarity between these two representations is measured using a similarity function, $\text{sim}(\cdot, \cdot)$, commonly implemented as cosine similarity. The temperature parameter τ is introduced to control the scale of the similarities, affecting the smoothness of the distribution.

The loss function's numerator, $\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)$, measures the similarity between the anchor and the positive sample, aiming to maximize it. The denominator, $\sum_{j=0}^{K} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)$, aggregates the similarities between the anchor and all samples in the batch, including both positive and negative samples.

We extract the tensors representing the predicted labels from the corresponding positions in M_{tl}, M_{tm}, M_a, M_v, and M_{fusion}, denoted as z_l, $\hat{z}_t$, $\hat{z}_v$, $\hat{z}_a$, and $\hat{z}_f$, respectively, to construct the contrastive learning loss. $\hat{z}_t$, $\hat{z}_v$, $\hat{z}_a$, and $\hat{z}_f$ represent the predictions of the label from the textual, visual acoustic, and multimodal fusion view, respectively. In contrast, z_l denotes the encoded representation of the true label. The final loss function of multi-view contrastive learning can be expressed as:

$$\mathcal{L}_{\text{text}} = \mathcal{L}_{\text{InfoNCE}}(\mathbf{z}_l, \widehat{\mathbf{z}}_t), \tag{23}$$

$$\mathcal{L}_{\text{visual}} = \mathcal{L}_{\text{InfoNCE}}(\mathbf{z}_l, \widehat{\mathbf{z}}_v), \tag{24}$$

$$\mathcal{L}_{\text{acoustic}} = \mathcal{L}_{\text{InfoNCE}}(\mathbf{z}_l, \widehat{\mathbf{z}}_a), \tag{25}$$

$$\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{InfoNCE}}(\mathbf{z}_l, \widehat{\mathbf{z}}_f), \tag{26}$$

$$\mathcal{L}_{\text{multi\_view}} = \frac{1}{4}(\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{visual}} + \mathcal{L}_{\text{acoustic}} + \mathcal{L}_{\text{fusion}}), \tag{27}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{multi\_view}} + \mathcal{L}_{\text{cross\_entropy}}. \tag{28}$$

Experiments

Datasets

We validate our proposed framework on two challenging datasets dedicated to intent recognition.

MIntRec MIntRec (Zhang et al. 2022a) dataset contains 2,224 high-quality samples and 20 intent labels. The samples cover text, visual, and acoustic modalities.

MIntRec2.0 MIntRec2.0 (Zhang et al. 2024) represents a large-scale benchmark dataset designed for multimodal intent recognition. This dataset encompasses 1,245 high-quality dialogues, aggregating a total of 15,040 samples that incorporate text, video, and audio modalities.

A distinguishing characteristic of the MIntRec 2.0 dataset is its fine-grained intent classification, featuring a robust categorization system with 30 distinct intent categories. Furthermore, the dataset is specifically engineered to include "out-of-range" samples that naturally arise in conversational exchanges, thereby providing a more rigorous testing scenario for evaluating model performance.

Baselines

In our experiments, we employ the following state-of-the-art method as baselines: (1) MulT (Tsai et al. 2019) is specifically designed for handling unaligned multimodal language sequences; (2) MAG-BERT (Rahman et al. 2020) leverages an attachment called the Multimodal Adaptation Gate (MAG) that seamlessly integrates with the pre-existing architecture of BERT without altering its fundamental structure; (3) TCL-MAP (Zhou et al. 2024) framework is specifically designed to enhance text modality representations by leveraging auxiliary video and audio modalities through a modality-aware prompting module (MAP), which aligns and fuses features across modalities using similarity-based modality alignment and cross-modality attention mechanisms. A more detailed description of the baselines can be found in the appendix.

Views				MIntRec				MIntRec 2.0			
Visual	Acoustic	Text	Fusion	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
✓	×	×	×	18.00	16.95	16.85	12.68	36.96	20.18	14.64	3.26
×	✓	×	×	25.39	22.80	23.47	18.15	37.06	20.04	13.73	3.23
×	×	✓	×	72.13	71.80	72.50	68.80	56.73	53.59	54.89	40.30
×	×	×	✓	74.45	74.39	**75.20**	**72.96**	57.64	54.88	**56.13**	41.53
✓	✓	✓	✓	**74.72**	**74.61**	75.07	71.94	**57.80**	**55.05**	55.82	**42.03**

Table 2: The results of ablation experiment for the MVCL-DAF method on MIntRec and MIntRec 2.0. This table shows the outcomes of the multi-view contrastive learning method when trained and inferred using each individual view, as well as the results obtained from training and inference using all views. The best performance for each metric is highlighted in bold.

Results

The experimental results are shown in Table 1, where we highlight the best value for each metric in bold. Delta represents the difference between the results of our method and the highest value of the corresponding evaluation metric among the baselines. Our method outperformed state-of-the-art methods across all metrics. On the MIntRec dataset, our method achieved ACC, WF1, WP, and R scores of 74.72%, 74.61%, 75.07%, and 71.94%, respectively, surpassing the state-of-the-art baseline by 1.03%, 1.23%, 1.17%, and 0.35%, respectively. On the MIntRec 2.0 dataset, even with an Unknown label accounting for as much as 38.28% of the samples, our method achieved significant improvements, reaching an ACC of 57.80%, WF1 of 55.05%, WP of 55.82%, and R of 42.03%. These results surpass the best baseline performance by 0.81%, 0.72%, 0.75%, and 0.16%, respectively. This observation strongly supports the effectiveness of the DAF module and the multi-view contrastive learning approach in the domain of multimodal intent recognition.

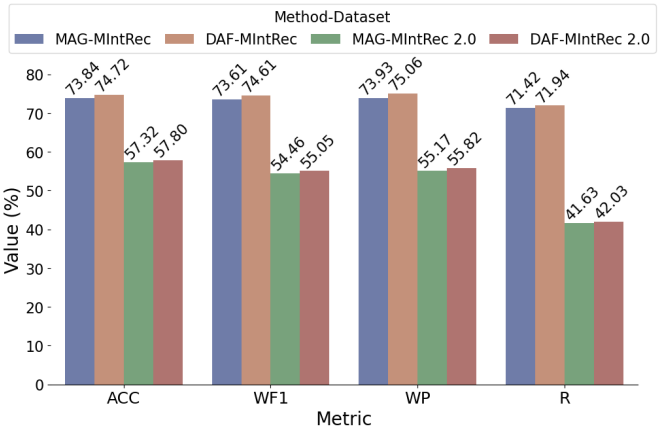

Figure 3: Performance of the multi-view contrastive learning method using the DAF fusion module compared to the MAG fusion module on the MIntRec and MIntRec 2.0 datasets.

Model Analysis

We did a number of ablation experiments to learn more about how the multi-view contrastive learning and DAF module affected the overall performance. Also, since both the MIntRec and MIntRec 2.0 datasets are meant to recognize intent and have some intent labels that overlap, we also conducted an experiment to validate the model's generalization capability.

Effectiveness of Dynamic Attention Allocation

Both MAG-BERT and TCL-MAP use MAG (Multimodal Adaptation Gate) as the multimodal fusion method, making the MAG module a suitable baseline for comparison. In our approach, we replaced the fusion module with MAG and conducted 10 experiments on the MIntRec and MIntRec 2.0 dataset using random seeds from 0 to 9. As shown in Figure 3, On the MIntRec dataset, the DAF module improved the model's performance by 0.88%, 1.00%, 1.13%, and 0.52% in ACC, WF1, WP, and R, respectively. On the MIntRec 2.0 dataset, the corresponding improvements were 0.48%, 0.60%, 0.65%, and 0.40%.

Figure 4 illustrates the distribution of attention scores for the visual and acoustic modalities across the two datasets. On the MIntRec dataset, the acoustic modality received higher attention scores, while the visual modality received less focus. However, within each modality, there is significant variation in attention distribution among samples. In contrast, on the MIntRec 2.0 dataset, the attention scores for the visual and acoustic modalities are much closer, with less variation between samples. This suggests that the DAF module is capable of not only broadly adjusting attention scores across samples and modalities but also fine-tuning the attention distribution within each modality.

Effectiveness of Multi-view Contrastive Learning We analyzed the impact of multi-view contrastive learning on model performance from two perspectives. First, we examined the effects of distributed versus centralized multimodal processing within our proposed method. Following this, we analyzed the impact of each modality and perspective on model performance.

Distributed Processing vs. Centralized Processing In our proposed method, the data from the three modalities are first aligned and then processed by their respective modality-specific encoders. Multimodal fusion occurs after each modality's information has been processed by its corresponding encoder. This is a distributed processing approach. In contrast, an alternative approach is to align and fuse the multimodal data first, and then process the fused data using a single centralized encoder. In our experiments, we used the bert-large-uncased model as the centralized encoder after fusion.

Methods	Training: MIntRec Testing: MIntRec 2.0				Training: MIntRec 2.0 Testing: MIntRec			
	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
MulT	61.86	61.25	63.66	59.58	70.22	69.45	72.78	69.92
MAG-BERT	**62.19**	61.42	63.00	59.73	71.06	69.77	71.81	68.63
TCL-MAP	61.97	61.74	63.52	59.96	70.25	68.64	70.52	68.22
MVCL-DAF	61.95	**61.89**	**63.68**	**60.65**	**71.68**	**70.81**	**73.49**	**70.81**

Table 3: The result of model generalization testing. The best result for each metric is highlighted in bold.

Our experimental results indicate that using a centralized encoder led to a performance decline. Specifically, ACC, WF1, WP, and R decreased by 0.72%, 0.39%, 0.78%, and 0.48%, respectively. These results demonstrate that distributed processing, which more closely mimics the way the human brain processes multimodal information, significantly improves model performance within our multi-view contrastive learning framework.

Impact of Views In our proposed method, we provide the model with four views: text, visual, acoustic, and a fusion perspective that combines the first three modalities. To analyze the contribution of each perspective to the model's performance, we conducted a series of single-perspective model training experiments and evaluated the models on the MIntRec and MIntRec 2.0 datasets. In this series of experiments, we only used the encoded high-dimensional vectors of the text modality and its corresponding labels to construct the anchor samples for contrastive learning. These high-dimensional vectors serve as a guide for learning from the other modalities. During model training and inference, positive samples for contrastive learning were constructed using data from a single perspective. Similarly, only the single perspective was used for representation learning. The experimental results, as shown in Tabel 2, indicate that the model struggles to accurately determine intent from the acoustic or visual perspectives alone. In contrast, the text perspective alone or the fused multimodal perspective provides more reliable intent predictions. Moreover, the performance of the model relying solely on the fusion view is already very close to that of the multi-view contrastive learning model, even surpassing the latter in some metrics. This clearly demonstrates the effectiveness of the dynamic attention allocation module.

Model Generalization Testing MIntRec is constructed using data collected from the TV series "Superstore." In contrast, MIntRec 2.0 incorporates data from three popular TV series: "Superstore," "The Big Bang Theory," and "Friends." This suggests that while the two datasets share some distributional similarities, they also exhibit noticeable differences. The label set of the MIntRec dataset is a subset of the label set in MIntRec 2.0. To align the label sets of the two datasets, we filtered the samples in MIntRec 2.0 to include only those with labels present in the MIntRec label set.

We designed two experimental scenarios to test generalization. In the first scenario, MIntRec was used as the training and validation dataset, with MIntRec 2.0 serving as the test dataset. In the second scenario, the roles of the two datasets were swapped. In each scenario, we trained and evaluated our proposed method and all baseline models using random seeds 0 to 4. The results, shown in Table 3, demonstrate that our proposed method exhibits strong generalization capabilities. When trained on the more complex dataset and tested on the relatively simpler dataset, MVCL-DAF shows significant advantages across all evaluation metrics.

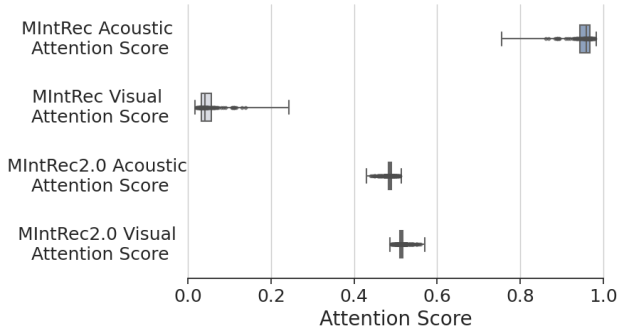

Figure 4: Distribution of attention scores for the visual and acoustic modalities on the MIntRec and MIntRec 2.0 datasets.

Conclusion

To address the issue of significant variation in the density of important information across different modalities and samples, we proposed a Dynamic Attention Allocation Fusion (DAF) method. This method utilizes an adaptive network structure to dynamically compute attention matrices and attention scores based on the input samples. The attention matrices assign higher weights to important information within each modality, while the attention scores dynamically adjust the overall weight distribution among the modalities. Building on this, we further introduced a multi-view contrastive learning method for intent recognition. Our method outperformed existing state-of-the-art models on the benchmark datasets MIntRec and MIntRec 2.0. Additionally, our ablation experiments confirmed that the text modality provides the primary basis for intent recognition. The fusion of multiple modalities and the application of multi-view contrastive learning effectively uncover important information from other modalities, leading to further improvements in model performance. Generalization experiments further revealed that models trained using the MVCL-DAF method exhibit excellent generalization capabilities, underscoring the robustness and effectiveness of our approach in diverse scenarios.

Acknowledgements

This research was supported by National Natural Science Foundation of China (No.62406303), Anhui Provincial Natural Science Foundation (No. 2308085QF229), Anhui Science and Technology Innovation Plan (No.202423k09020010) and the Fundamental Research Funds for the Central Universities (No. WK2150110034).

References

Aneja, J.; Agrawal, H.; Batra, D.; and Schwing, A. 2019. Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chong, R.; Kong, C.; Wu, L.; Liu, Z.; Jin, Z.; Yang, L.; Fan, Y.; Fan, H.; and Yang, E. 2023. Leveraging Prefix Transfer for Multi-Intent Text Revision. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1219–1228. Toronto, Canada: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gers, F. A.; and Schmidhuber, J. 2000. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, 189–194. IEEE.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 369–376. New York, NY, USA: Association for Computing Machinery. ISBN 1595933832.

Grill-Spector, K.; and Malach, R. 2004. The human visual cortex. *Annu. Rev. Neurosci.*, 27(1): 649–677.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. arXiv:2109.00412.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Joo, J.; Li, W.; Steen, F. F.; and Zhu, S.-C. 2014. Visual Persuasion: Inferring Communicative Intents of Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaas, J.; O'Brien, B. M.; and Hackett, T. A. 2012. Auditory processing in primate brains. *Handbook of Psychology, Behavioral Neuroscience*, 3: 157.

Kaas, J. H.; and Hackett, T. A. 2000. Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, 97(22): 11793–11799.

Liang, P. P.; Deng, Z.; Ma, M.; Zou, J.; Morency, L.-P.; and Salakhutdinov, R. 2023. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. arXiv:2306.05268.

Ni, J.; Bai, Y.; Zhang, W.; Yao, T.; Yang, Q.; Mei, T.; and Han, K. 2024. Deep Equilibrium Multimodal Fusion.

Ouyang, Y.; Ye, J.; Chen, Y.; Dai, X.; Huang, S.; and Chen, J. 2021. Energy-based unknown intent detection with data manipulation. *arXiv preprint arXiv:2107.12542*.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Romo, R.; and Salinas, E. 2003. Flutter discrimination: neural codes, perception, memory and decision making. *Nature Reviews Neuroscience*, 4(3): 203–218.

Stein, B. E.; and Meredith, M. A. 1993. *The merging of the senses*. MIT press.

Sun, K.; Xie, Z.; Ye, M.; and Zhang, H. 2024. Contextual Augmented Global Contrast for Multimodal Intent Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26963–26973.

Tang, X.; Liu, K.; Cui, J.; Wen, F.; and Wang, X. 2012. IntentSearch: Capturing User Intention for One-Click Internet Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7): 1342–1353.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Wandell, B. A.; and Winawer, J. 2015. Computational neuroimaging and population receptive fields. *Trends in cognitive sciences*, 19(6): 349–357.

Xue, Z.; and Marculescu, R. 2023. Dynamic multimodal fusion. In *Multi-Modal Learning and Applications Workshop (MULA), CVPR*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. Copenhagen, Denmark: Association for Computational Linguistics.

Zhang, H.; Li, X.; Xu, H.; Zhang, P.; Zhao, K.; and Gao, K. 2021a. TEXTOIR: An Integrated and Visualized Platform for Text Open Intent Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, 167–174.

Zhang, H.; Wang, X.; Xu, H.; Zhou, Q.; Gao, K.; Su, J.; jinyue Zhao; Li, W.; and Chen, Y. 2024. MIntRec2.0: A Large-scale Benchmark Dataset for Multimodal Intent Recognition and Out-of-scope Detection in Conversations. In *The Twelfth International Conference on Learning Representations*.

Zhang, H.; Xu, H.; Wang, X.; Zhou, Q.; Zhao, S.; and Teng, J. 2022a. MIntRec: A New Dataset for Multimodal Intent Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1688–1697. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Zhang, K.; Liu, Q.; Qian, H.; Xiang, B.; Cui, Q.; Zhou, J.; and Chen, E. 2021b. EATN: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 377–389.

Zhang, K.; Zhang, H.; Liu, Q.; Zhao, H.; Zhu, H.; and Chen, E. 2019. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5773–5780.

Zhang, K.; Zhang, K.; Zhang, M.; Zhao, H.; Liu, Q.; Wu, W.; and Chen, E. 2022b. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2203.16369*.

Zhou, Q.; Xu, H.; Li, H.; Zhang, H.; Zhang, X.; Wang, Y.; and Gao, K. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17114–17122.

Zou, D.; Wei, W.; Mao, X.-L.; Wang, Z.; Qiu, M.; Zhu, F.; and Cao, X. 2022a. Multi-level Cross-view Contrastive Learning for Knowledge-aware Recommender System. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

Zou, Y.; Liu, H.; Gui, T.; Wang, J.; Zhang, Q.; Tang, M.; Li, H.; and Wang, D. 2022b. Divide and Conquer: Text Semantic Matching with Disentangled Keywords and Intents. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 3622–3632. Dublin, Ireland: Association for Computational Linguistics.