WDMIR: Wavelet-Driven Multimodal Intent Recognition

Weiyin Gong[1,2], **Kai Zhang**[1,*], **Yanghai Zhang**[1], **Qi Liu**[1], **Xinjie Sun**[1,2], **Junyu Lu**[3] and **Linbo Zhu**[3]
[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]School of Computer Science, Liupanshui Normal University
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{weiyingong,yhzhang0612, xinjiesun,lujunyu}@mail.ustc.edu.cn, {kkzhang08,qiliuql}@ustc.edu.cn,
lbzhu@iai.ustc.edu.cn

Abstract

Multimodal intent recognition (MIR) seeks to accurately interpret user intentions by integrating verbal and non-verbal information across video, audio and text modalities. While existing approaches prioritize text analysis, they often overlook the rich semantic content embedded in non-verbal cues. This paper presents a novel *Wavelet-Driven Multimodal Intent Recognition* ***(WDMIR)*** framework that enhances intent understanding through frequency-domain analysis of non-verbal information. To be more specific, we propose: (1) a wavelet-driven fusion module that performs synchronized decomposition and integration of video-audio features in the frequency domain, enabling fine-grained analysis of temporal dynamics; (2) a cross-modal interaction mechanism that facilitates progressive feature enhancement from bimodal to trimodal integration, effectively bridging the semantic gap between verbal and non-verbal information. Extensive experiments on MIntRec demonstrate that our approach achieves state-of-the-art performance, surpassing previous methods by 1.13% on accuracy. Ablation studies further verify that the wavelet-driven fusion module significantly improves the extraction of semantic information from non-verbal sources, with a 0.41% increase in recognition accuracy when analyzing subtle emotional cues.

1 Introduction

Intent recognition is a key aspect of human-computer interaction, and its core goal is to enable machines to accurately grasp user intent and thus provide users with better service [Zhang *et al.*, 2019; Huang *et al.*, 2023; Qiu, 2024]. Recently, multimodal intent recognition has been used to understand the user's intent in more complex scenarios. Compared with unimodal intent, multimodal information fusion can improve intent recognition accuracy by making joint decisions [Soleymani *et al.*, 2017; Chen *et al.*, 2021; Zou *et al.*, 2022; Zhang *et al.*, 2024; Sun *et al.*, 2025].

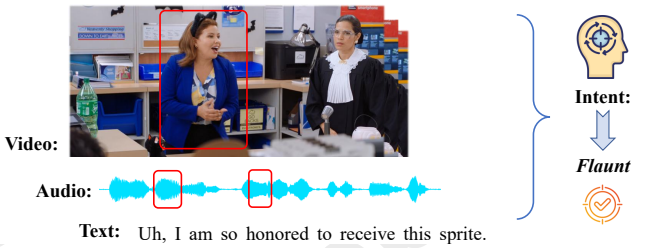


Figure 1: In the task of multimodal intent recognition, videos and audio contain relatively few key pieces of information, necessitating the in-depth mining of complementary information.

Researchers are currently paying greater attention to multimodal intent detection research as it focuses more on intricate real-world situations. To this end, [Zhang *et al.*, 2022a] introduces the first multimodal intention recognition baseline dataset, MIntRec, and conducts experiments on models such as MulT [Tsai *et al.*, 2019], MISA [Hazarika *et al.*, 2020], and MAG-BERT [Rahman *et al.*, 2020] to establish a baseline for intent recognition for subsequent research. Subsequently, in order to effectively integrate information from different modalities such as text, video, and audio, researchers have designed various models, all of which have achieved certain results in the field of multimodal intent recognition [Zhang *et al.*, 2022c; Sun *et al.*, 2024; Huang *et al.*, 2024; Zhou *et al.*, 2024]. Although existing methods have made some progress in multimodal intent recognition, the accuracy of multimodal analysis of user's intent is still limited due to the fact that the potential correlation between different modalities has not yet been fully explored, as well as the insufficiency of the video and audio modalities in semantic feature extraction. As shown in Figure 1, when the speaker says, "Uh, I am so honored to receive this sprite," we might initially infer that the speaker is joking. However, by carefully observing the speaker's facial expressions and analyzing their tone, we can discern that the speaker's true intention is to flaunt. so much so that multimodal intention recognition currently presents the following two challenges: first, how to deeply mine the semantic information in video and audio modalities; and second, how to efficiently align and fuse the features of text, video and audio modalities.

To address the first challenge, we propose a wavelet-driven approach. As far as we know, we are the first to

*Corresponding author.

introduce wavelet transform to drive the fusion of video and audio data information. Wavelet transform decomposes signals into low-frequency and high-frequency components. The low-frequency component captures the global characteristics of the signal, reflecting smooth trends and large-scale variations [Satirapod *et al.*, 2001], while the high-frequency component focuses on local details, such as abrupt changes, fine structures, and rapid variations [Lahmiri, 2014; Li *et al.*, 2023].

To address the second challenge, we designed collaborative representations and progressive fusion modules. These modules aim to enhance the alignment and integration between wavelet-driven non-verbal modalities and text modalities through cross-modal mechanisms, achieving a transition from bimodal to trimodal collaborative representations. Subsequently, the progressive fusion module is utilized for deep representation, leveraging complementary information at different stages to improve the accuracy of multimodal intent recognition. Our primary contributions in this study are:

- We design wavelet-driven multimodal intent recognition methods to achieve fusion of nonverbal modal features in the frequency domain to improve the model's ability to understand and recognize multimodal data.
- We achieve cross-modal collaborative alignment to integrate multimodal information, and through progressive fusion, mine complementary information to enhance recognition accuracy.
- Our experiments significantly improved each metric on MIntRec and MELD-DA, validating the validity and generalizability of the method.

2 Related Works

2.1 Multimodal Intent Recognition

Multimodal intent recognition aims to extract the user intent from multimodal information. However, challenges remain when dealing with scenarios involving text, visual, and acoustic modalities. [Zhang *et al.*, 2022a] introduced the multimodal intent recognition task and released a MIntRec dataset, which integrates textual, visual, and acoustic modalities to recognize user intent comprehensively. [Zhou *et al.*, 2024] proposed a token-level contrastive learning method with modality-aware prompts that effectively integrates text, audio, and video modality features through similarity-based modality alignment and cross-modal attention. [Huang *et al.*, 2024] introduced a shallow-to-deep interactive framework with data augmentation capabilities to address the modality alignment issue by gradually fusing multimodal features and incorporating ChatGPT-based data augmentation methods. [Sun *et al.*, 2024] proposed a context-enhanced global contrastive method that alleviates biases and inconsistencies in multimodal intent recognition by using within-video and cross-video interactions and retrieval, combined with global context-guided contrastive learning.

2.2 Multimodal Fusion Methods

Multimodal fusion aims to integrate information from different modalities (e.g., text, video, and audio) to better understand and process information from multiple sources, thereby improving system performance [Zhang *et al.*, 2022b]. For example, [Zadeh *et al.*, 2017] proposes a tensor fusion network that learns intra- and inter-modal dynamic features in an end-to-end manner. [Tsai *et al.*, 2019] addresses the problem of aligning multimodal sequences in different time steps by using directed pairwise cross-modal attention. [Rahman *et al.*, 2020] Introduced multimodal adaptation gates to fine-tune BERT to address non-verbal modal input. [Hazarika *et al.*, 2020] maps each modality to two different subspaces to learn common and modality-specific features.

2.3 Wavelet Transform

Wavelet transform is a commonly used time-frequency analysis technique in signal processing that can convert information from the time domain to the frequency domain. The decomposed information exhibits both global and local characteristics. In recent years, many researchers have applied wavelet transforms to various fields. [Li *et al.*, 2023] introduced a wavelet fusion module for facial super-resolution tasks, addressing the issue in existing Transformer-based methods where global information integration often neglects relevant details, leading to blurring and limiting high-frequency detail recovery. [Phutke *et al.*, 2023] proposed an end-to-end blind image inpainting architecture with a wavelet query multi-head attention transformer block, effectively repairing images by using the wavelet coefficients processed to provide encoder features as queries, bypassing the damaged region prediction step. [Sabry *et al.*, 2024] applies wavelet transform to lung sound signal analysis, addressing the issues of noise contamination and artifact removal in lung sound signals. By performing a multi-scale analysis of the lung sound signals, effective features are extracted, improving the accuracy of lung disease classification. [Frusque and Fink, 2024] introduced wavelet techniques for denoising audio information. Our method mainly uses wavelet transform to decompose video and audio information at multiple levels, to realize the fusion of video and audio features in the frequency domain, and to realize the analysis of non-verbal modal information in the frequency domain to enhance its representation.

3 Method

3.1 Task Description

Let text, video, and audio data from three modalities are taken as inputs, their features are fused to predict user intent. Specifically, let the feature of the text modality be t, the feature of the video modality be v, and the feature of the audio modality be a. By designing a multimodal fusion module $\mathcal{F}$, these features are mapped into a unified representation.

$$\mathbf{H} = \mathcal{F}(t, v, a) \tag{1}$$

Subsequently, a classifier $\mathcal{F}_{\text{intent}}$ maps the fused features to the intent category space, generating a predicted distribution

$$\hat{y} = \mathcal{F}_{\text{intent}}(\mathbf{H}) \tag{2}$$

where $t \in \mathbb{R}^{d_t}$, $v \in \mathbb{R}^{d_v}$, and $a \in \mathbb{R}^{d_a}$ represent the features of the text, video, and audio modalities, respectively, with d_t, d_v, and d_a denoting the dimensions of each modality's features. $\mathbf{H} \in \mathbb{R}^{d_h}$ is the fused feature representation; d_h is the dimension of the fused features; $\hat{y} \in \mathbb{R}^c$ represents the predicted probability distribution over c intent categories.

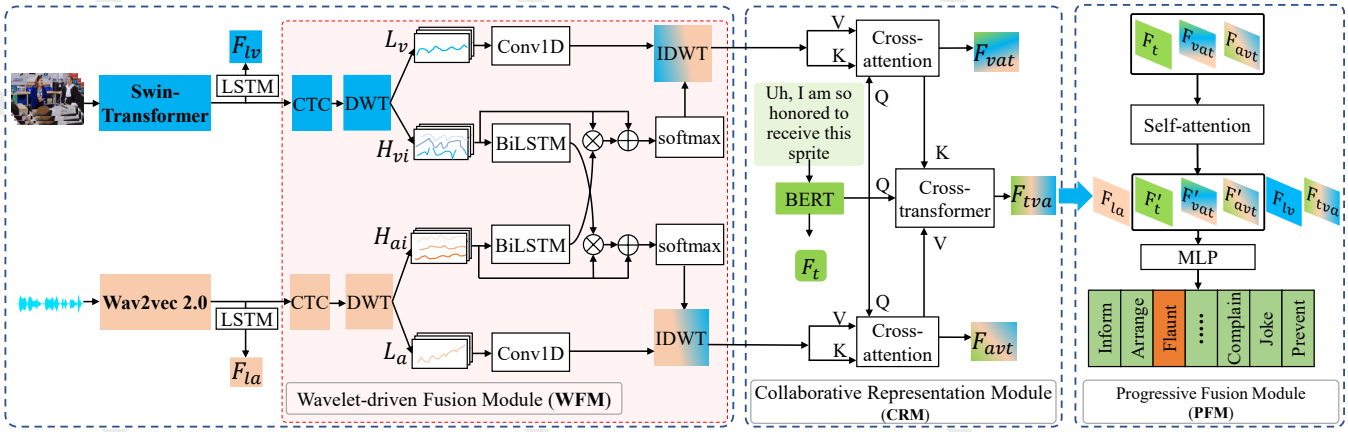


Figure 2: The overall framework of WDMIR primaril consists of the Wavelet-driven Fusion Module (WFM), Collaborative Representation Module (CRM), and Progressive Fusion Module (PFM), which are responsible for video and audio feature fusion, cross-modal collaborative representation, and enhancing intent recognition accuracy using multi-level features, respectively.

3.2 Framework Overview

The overall framework of WDMIR is shown in Figure 2. It mainly consists of Wavelet-driven Fusion Module (WFM), Collaborative Representation Module (CRM) and Progressive Fusion Module (PFM). WFM can realize the feature fusion of video and audio in frequency domain by multilevel decomposition of video and audio in the frequency domain. CRM utilizes the cross-modal mechanism to realize the collaborative representation of text modality and wavelet-driven video and audio modality to realize the collaborative representation of two modalities to three modalities. PFM improves the accuracy of intent recognition using multilevel feature progressive fusion representation.

3.3 Feature Extraction

For text modality, in order to fully utilize the semantic information in the text [Liu *et al.*, 2023], we use pre-trained model BERT [Lee and Toutanova, 2018] to encode the text and extract the last hidden layer as the feature representation of the text.

$$F_t = \text{BERTEmbeding}(t) \tag{3}$$

where t is the input conversation text, F_t is the text feature extracted through BERTEmbedding.

For the video modality, we follow the research approach in [Zhou *et al.*, 2024] and choose the Swin-Transformer [Liu *et al.*, 2021], which performs excellently in the field of computer vision, to sample the video frame by frame and perform pre-training, extracting the last hidden layer to represent the visual semantic information.

$$F_v = \text{Swin-Transformer}([f_1, f_2, .., f_{lv}]) \tag{4}$$

where f_i is the i-th frame in each video segment, lv is the number of frames sampled from the video, and F_v is the visual semantic information extracted by Swin-Transformer.

For audio modality, we adopt the method [Zhang *et al.*, 2022a] to pre-train the model using Wav2Vec 2.0 [Baevski *et al.*, 2020] to extract the output of the last hidden layer as the audio feature representation.

$$F_a = \text{Wav2Vec 2.0}(a) \tag{5}$$

where a is the audio sequence corresponding to the video clip and F_a is the audio feature extracted by Wav2Vec 2.0.

3.4 Wavelet-driven Fusion Module (WFM)

WFM is able to decompose audio and video signals over sequences into the frequency domain, where fine-grained feature fusion of non-verbal modalities is achieved.Before further analyzing the video and audio features, we perform sequence alignment of the video and audio features using the CTC model [Graves *et al.*, 2006].

$$V, A = CTC(F_v, F_a) \tag{6}$$

where V and A represent the video and audio features after sequence alignment, respectively.

In order to facilitate the fusion of non-verbal modal information in the frequency domain. We choose Haar wavelet bases to perform 3-level 1-dimensional DWT transform on audio and video aligned in sequences to obtain the high-frequency and low-frequency components of video and audio information, which correspond to local and global features, respectively.

$$\begin{aligned}(L_a, H_{ai}) &= DWT(A)\\ (L_v, H_{vi}) &= DWT(V)\end{aligned} \tag{7}$$

Where L_a and L_v are the low-frequency components of the audio and video features obtained through multilevel wavelet decomposition, respectively, and H_{ai} and H_{vi} are the i-th high-frequency components of the audio and video features after multilevel decomposition, respectively.

To better extract global information from video and audio data, we employ a one-dimensional convolution with a kernel size of 3 to strengthen the representation of low-frequency components, further enhancing the model's understanding of global information.

$$\begin{aligned}L_a^{'} &= Conv1D(L_a)\\ L_v^{'} &= Conv1D(L_v)\end{aligned} \tag{8}$$

where $L_a^{'}$ and $L_v^{'}$ are the low-frequency components of audio and video enhanced by one-dimensional convolution.

For the multilevel high-frequency components obtained from wavelet decomposition, we concatenate the components at each level sequentially to construct the total high-frequency components for each modality. This approach integrates high-frequency information from different levels, enhancing the model's ability to capture fine-grained features.

$$\begin{aligned} F_{hv} &= cat(H_{v1}, H_{v2}, H_{v3}) \\ F_{ha} &= cat(H_{a1}, H_{a2}, H_{a3}) \end{aligned} \tag{9}$$

where F_{hv} and F_{ha} are the total high-frequency components of video and audio, respectively.

To better facilitate the fusion of video and audio information, we use the BiLSTM model to map audio to the video space and video to the audio space, respectively.

$$\begin{aligned} H_v^{'} &= BiLSTM(F_{hv}) \\ H_a^{'} &= BiLSTM(F_{ha}) \end{aligned} \tag{10}$$

where $H_v^{'}$ is the visual feature obtained from the high-frequency component of the video through BiLSTM; $H_a^{'}$ is the audio feature obtained from the high-frequency component of the audio through BiLSTM.

To better leverage the complementary information between modalities and enhance feature representation, we designed a frequency domain interaction module.

$$\begin{aligned} H_{AV} &= softmax(F_{ha} \odot H_v^{'} + F_{ha}) \\ H_{VA} &= softmax(F_{hv} \odot H_a^{'} + F_{hv}) \end{aligned} \tag{11}$$

where H_{AV} is the audio feature obtained from the interaction between audio and video, H_{VA} is the video feature obtained from the interaction between video and audio, and $\odot$ denotes the element-wise multiplication.

To better integrate the interaction-enhanced features with their respective low-frequency features, we perform an inverse wavelet transform to reconstruct the audio and video modalities, resulting in fused audio and video features.

$$\begin{aligned} F_{AV} &= IDWT(cat(L_a^{'}, H_{AV})) \\ F_{VA} &= IDWT(cat(L_v^{'}, H_{VA})) \end{aligned} \tag{12}$$

where F_{AV} is the reconstructed and enhanced audio modality feature, and F_{VA} is the reconstructed video modality feature.

3.5 Collaborative Representation Module (CRM)

The text modality serves as the primary source of information for intent recognition. We treat the text modality as the main modality and perform pairwise interactions with the enhanced audio and video modalities separately. Cross-modal attention mechanisms are employed to achieve alignment and feature fusion. Subsequently, the audio and video modalities, weighted by the text features, are further interacted with the text modality, enabling deep fusion of the three modalities.

Considering that the text modality contains the primary information, we perform sequence alignment on the fused video and audio features through a cross-attention mechanism [Huang *et al.*, 2024]. We choose the text modality F_t as the query Q, and the fused video F_{VA} and audio modality F_{AV} as the key and value K and V.

$$\begin{aligned} Cross-attention(Q,K,V) &= softmax(\frac{QK^T}{\sqrt{d_k}})V \\ F_{vat} &= Cross-attention(F_t, F_{VA}, F_{VA}) \\ F_{avt} &= Cross-attention(F_t, F_{AV}, F_{AV}) \end{aligned} \tag{13}$$

where F_{vat} and F_{avt} denote the weighted video and audio features, respectively.

Then F_t as the query vector Q, F_{vat} as the key vector K, and F_{avt} as the value vector V are used to realize the trimodal co-representation via cross-transformer [Tsai *et al.*, 2019].

$$F_{tva} = softmax(\frac{F_t {F_{vat}}^T}{\sqrt{d_k}})F_{avt} \tag{14}$$

where F_{tva} denotes the three modal co-representation features of text, video, and audio.

3.6 Progressive Fusion Module (PFM)

PFM improves the accuracy of multimodal intent recognition mainly through multilevel feature progressive fusion.To focus on the fused text, video, and audio model features from different visions. We stack F_t, F_{vat} and F_{avt} to obtain the matrix F_m.

$$F_m = [F_t, F_{vat}, F_{avt}] \tag{15}$$

Then, the stacked F_m is enhanced through self-attention [Vaswani, 2017; Zhang *et al.*, 2021] for feature enhancement.

$$F = Self-attention(F_m) \tag{16}$$

where $F = [F_t^{'}, F_{vat}^{'}, F_{avt}^{'}]$ denotes the enhanced features.

We utilize LSTM to learn the encoded video and audio information, taking the hidden state of the last layer as output to prevent the loss of key features during the processing of video and audio characteristics, thereby enhancing the model's understanding of complex data.

$$\begin{aligned} F_{lv} &= LSTM(F_v) \\ F_{la} &= LSTM(F_a) \end{aligned} \tag{17}$$

where F_{lv} and F_{la} denote the video features and audio features obtained by LSTM, respectively.

We concatenate the features obtained from multiple layers and map them to the output space through an MLP to get the final output $\hat{y}$.

$$\hat{y} = MLP(cat(F_{lv}, F_{la}, F_t^{'}, F_{vat}^{'}, F_{avt}^{'}, F_{tva})) \tag{18}$$

To make the output better approximate the true distribution, we use a cross-entropy loss function to measure the difference between the predicted and true values.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \tag{19}$$

where $\mathcal{L}$ is the total loss function of the model. N is the number of samples and C is the number of intended categories. y_{ij} is the true label of the i-th sample in the jth category. $\hat{y}_{ij}$ is the probability distribution of the i-th sample belonging to the j-th category as predicted by the model.

Methods	MIntRec				MELD-DA			
	ACC	**WF1**	**WP**	**R**	**ACC**	**WF1**	**WP**	**R**
MAG-BERT	72.65	72.16	72.53	69.28	60.63	59.36	59.80	50.01
MISA	72.29	72.38	73.48	69.24	59.98	58.52	59.28	48.75
MulT	72.52	72.31	72.85	69.24	60.36	59.01	59.44	49.93
TCL-MAP	73.62	73.31	73.72	70.50	<u>61.75</u>	<u>59.77</u>	<u>60.33</u>	<u>50.14</u>
SDIF-DA∗	<u>73.93</u>	<u>73.89</u>	<u>74.18</u>	<u>71.66</u>	−−	−−	−−	−−
CAGC	73.39	−−	−−	70.39	−−	−−	−−	−−
WDMIR (Our)	**75.06**	**74.96**	**75.26**	**72.65**	**62.16**	**60.94**	**61.44**	**51.36**
△	1.13↑	1.07↑	1.08↑	0.99↑	0.41↑	1.17↑	1.11↑	1.22↑

Table 1: The experimental results of our method on the MIntRec and MELD-DA datasets are as follows. △ represent the comparison results between our method and the previous best method, bold indicates the best results, underline denotes the second best results, ↑ denotes the improvement effect, asterisks * indicate the results from our re-experimentation, and all other results are sourced from published papers.

4 Experiments

4.1 Datasets

We conduct experiments on two datasets, MIntRec [Zhang *et al.*, 2022a] and MELD-DA [Saha *et al.*, 2020]. MIntRec is a multimodal intent dataset containing text, video, and audio, with 2224 samples and 20 intent categories. It includes 1334, 445, and 445 samples for training, validation, and testing, respectively. MELD-DA is a multi-round emotion conversation dataset containing text, video, and audio, with 9988 samples and 12 emotion conversation behavior labels. It includes 6991, 999, and 1998 samples for training, validation, and testing, respectively.

4.2 Baselines

In our experiments, we will use state-of-the-art multimodal fusion methods as baselines: (1) MISA [Hazarika *et al.*, 2020] projects each modality into two different subspaces to learn the fusion of common features and unique attributes of different modalities. (2) MulT [Tsai *et al.*, 2019] potentially converts one modality to another for feature fusion by directing paired cross-modal attention. (3) MAG-BERT [Rahman *et al.*, 2020] introduces a multimodal adaptation gate as an accessory to fine-tune BERT to enable it to handle non-verbal data for multimodal data fusion. (4) TCL-MAP [Zhou *et al.*, 2024] uses the similarity of modalities to design a multimodal perceptual cueing module for modal alignment, and uses a cross-modal attention mechanism to generate modal perceptual cues for multimodal fusion. (5) SDIF-DA [Huang *et al.*, 2024] enhances text data through ChatGPT, then designs interaction modules from shallow to deep and gradually aligns and fuses features between different modalities effectively. (6) CAGC [Sun *et al.*, 2024] enhances the capture of global contextual features by mining the contextual interaction information within and between videos, thus effectively solving the problems of perceptual bias and inconsistency of multimodal representations.

4.3 Evaluation Metrics

Based on previous work [Zhou *et al.*, 2024], we use Accuracy (ACC), Weighted F1 Score (WF1), Weighted Precision (WP), and Recall (R) as the evaluation metrics of the model. The impact of sample unevenness on model performance is reduced by a weighted average of the number of samples in each category by WF1 and WP.

4.4 Experimental Settings

In our experiments, bert-base-uncased[1] and wav2vec2-base-960h[2] from the Huggingface are used as the pre-training models for extracting mentioned text and audio features. Video features are extracted from the Torchvision library by using swin_b pre-trained on ImageNet1K. Adam [Loshchilov, 2017] as an optimization parameter throughout the experiment. The training batch size is 16, and the validation and test batch sizes are both 8.

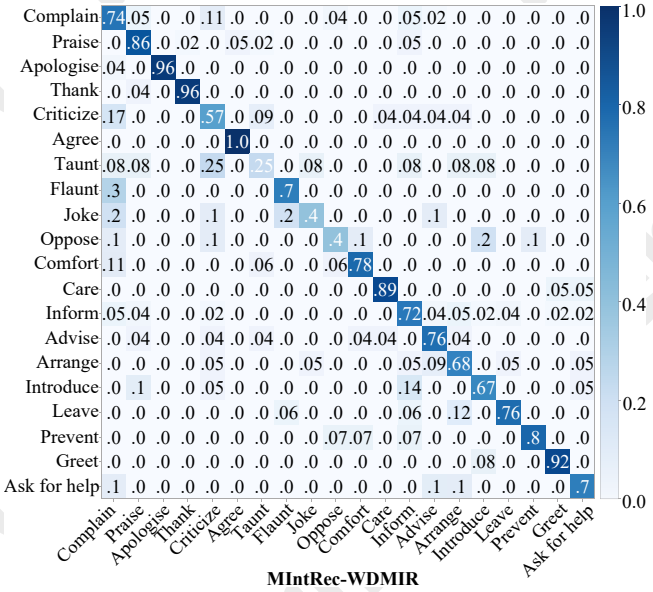
4.5 Main Result

Our method is compared with the optimal method in §4.2, and the experimental results are shown in Table 1. From the analysis of the experimental results, it is clear that our approach has made significant progress in two main areas. First, on the MIntRec dataset, our method achieves 1.13%, 1.07%, 1.08%, and 0.99% improvement in four key performance metrics, namely, ACC, WF1, WP, and R, respectively, when compared to the best existing baseline method. These results strongly demonstrate the effectiveness of the method in handling multimodal intent recognition tasks in complex real-world scenarios. Second, on the multi-round sentiment dialog analysis dataset MELD-DA, our method also demonstrates better performance, improving 0.41%, 1.17%, 1.11%, and 1.22% in the four evaluation metrics of ACC, WF1, WP, and R, respectively, compared to the optimal baseline. This result confirms the effectiveness and robustness of our proposed method in the task of conversational emotion recognition. Overall, the experimental results fully demonstrate that

[1]https://huggingface.co/google-bert/bert-base-uncased
[2]https://huggingface.co/facebook/wav2vec2-base-960h

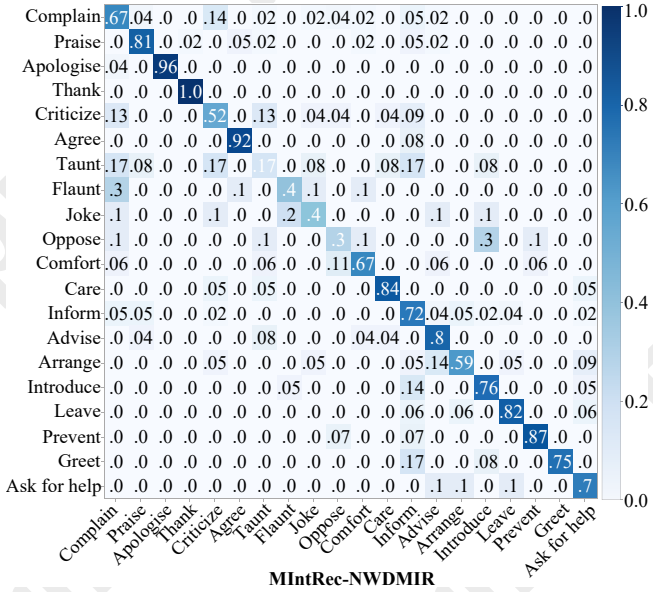
Modules						MIntRec				MELD-DA			
F_{lv}	F_{la}	$F^{'}_{vat}$	$F^{'}_{avt}$	F_{tva}	WFM	ACC	WF1	WP	R	ACC	WF1	WP	R
×	×	✓	✓	✓	✓	73.49	73.24	73.46	70.68	61.68	60.56	60.65	51.20
✓	✓	×	×	✓	✓	72.70	72.29	72.36	69.67	60.61	59.15	59.38	49.59
✓	✓	✓	✓	×	✓	72.02	71.71	71.82	69.06	61.06	59.63	59.66	50.61
✓	✓	✓	✓	✓	×	72.02	71.76	72.20	68.77	60.56	59.28	59.13	49.68
✓	✓	✓	✓	✓	✓	**75.05**	**74.96**	**75.26**	**72.65**	**62.16**	**60.94**	**61.44**	**51.36**

Table 2: Conducting ablation studies on the MIntRec and MELD-DA datasets respectively. Features obtained from the first layer are denoted as F_{lv} and F_{la}, features from the second layer are denoted as $F^{'}_{vat}$ and $F^{'}_{avt}$, and features from the third layer are denoted as F_{tva}. WFM stands for the audio-video feature fusion module.



MIntRec-WDMIR

Figure 3: MIntRec-WDMIR represents the confusion matrix evaluation results with wavelet-driven fusion module on the MIntRec dataset, where leading zeros before decimal points are omitted.



MIntRec-NWDMIR

Figure 4: MIntRec-NWDMIR denotes the confusion matrix evaluation results obtained after removing the wavelet-driven fusion module in the MIntRec dataset, with leading zeros before decimal points omitted.

our method outperforms the existing state-of-the-art baseline methods on all evaluation metrics for two representative datasets, which in turn validates the effectiveness and generalizability of the method.

4.6 Ablation Study

To further explore the impact of different modules on the performance of the WDMIR method on the MIntRec, we perform ablation experiments on the Wavelet-driven Fusion Module, Collaborative Representation Module, and Progressive Fusion Module to determine the contribution of each module to the overall performance.

Wavelet-driven Fusion Module

To evaluate the performance of WFM, we removed it from the model. The experimental results are shown in Table 2: on the MIntRec dataset, the model's accuracy (ACC) dropped by 3.03%, weighted F1 score (WF1) decreased by 3.2%, weighted precision (WP) declined by 3.06%, and recall (R) decreased by 3.88%. On the MELD-DA dataset, ACC dropped by 1.6%, WF1 by 1.66%, WP by 2.31%, and R by 1.68%. These results indicate that WFM can effectively extract fine-grained semantic information from non-verbal modalities in intent and emotion data. Moreover, the confusion matrix in Figure 3 and the visualization in Figure 4 on the MIntRec dataset suggest that WFM also contributes to the recognition of implicit intents.

Collaborative Representation Module

To assess the effectiveness of the collaborative representation module, we conducted an ablation study by removing F_{tva}, with the results shown in Table 2. The removal of F_{tva} led to a significant performance drop on both the MIntRec and MELD-DA datasets, particularly on MIntRec, where ACC, WF1, WP, and R decreased by 3.03%, 3.25%, 3.44%, and 3.59%, respectively. These results demonstrate that the collaborative representation module effectively integrates complementary information across multiple modalities, thereby

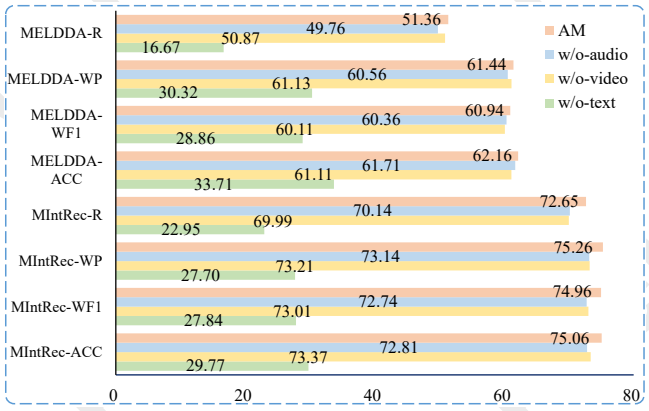


Figure 5: Single-Modality Missing Comparison. AM represents no missing modality, w/o-text indicates the removal of text modality, w/o-video indicates the removal of video modality, and w/o-audio indicates the removal of audio modality.

enhancing the accuracy and robustness of intent recognition.

Progressive Fusion Module

To verify the impact of the progressive module on the overall performance of WDMIR, we sequentially removed F_{lv} and F_{la}, as well as F'_{vat} and F'_{avt}. The experimental results are shown in Table 2. The results demonstrate that the progressive fusion module improves model performance on both the MIntRec and MELD-DA datasets. Specifically, F_{lv} and F_{la} compensate for the loss of audio-visual information, while F'_{vat} and F'_{avt} enhance deep fusion between modalities. The ablation experiments further validate the critical role of the progressive fusion module in optimizing multimodal information complementarity and improving the overall performance of the model.

4.7 Performance with Single-Modality Missing

We further analyzed the impact of removing a specific modality on the model's performance, as shown in Figure 5. The vertical axis of the figure depicts the model's performance metrics—ACC, WF1, WP, and R—on the MIntRec and MELD-DA datasets. The experimental results indicate that the absence of the text modality leads to a significant decline in performance, demonstrating that text plays a dominant role in multimodal tasks. Removing the video or audio modalities also causes performance degradation, but to a lesser extent, suggesting that they mainly provide auxiliary information.

4.8 F1-Score Analysis Across Intent Categories

To better evaluate the performance of WDMIR on MIntRec, we compare its f1-score with baselines to assess how each method performs across fine-grained intent categories. As illustrated in Figure 6, our approach demonstrates strong performance, particularly in the intent categories of "Flaunt", "Joke" and "Oppose" suggesting that WDMIR is adept at managing complex intent categories. WDMIR performs similarly to other methods in the "Care" and "Complain" categories, and even outperforms in some areas, indicating stable performance in common intent categories, likely due to

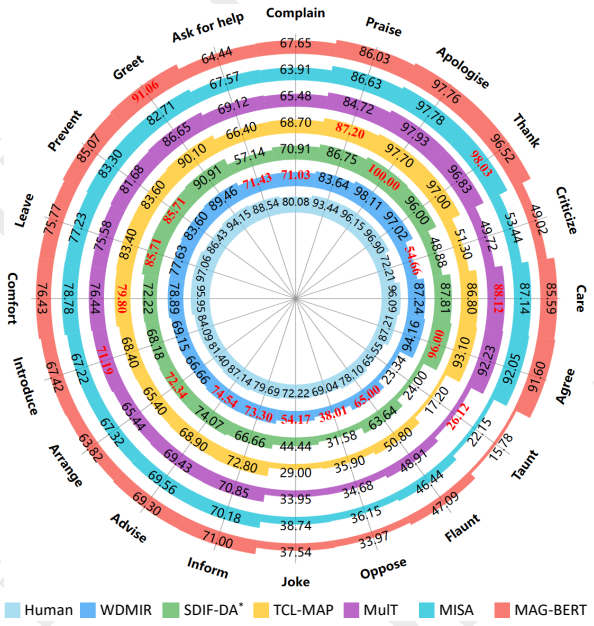


Figure 6: F1-score comparison between our method and the baselines on the MIntRec dataset. SDIF-DA* denotes the re-implemented result, while the other data are from TCL-MAP. Bold indicates the best performance excluding human results.

the wavelet-driven nonverbal modal fusion. However, despite some improvements in certain intent categories, WDMIR's performance in categories like "Taunt", "Oppose" and "Joke" is not as strong as that of other methods, which may be linked to the influence of wavelet-driven nonverbal modal fusion. The F1-scores metric performed significantly below human levels on the "Taunt", "Oppose", and "Joke" intentions. This suggests that WDMIR still faces challenges in understanding emotions that are heavily dependent on context and humor, and that further improvements in understanding linguistic humor and contextual variation are needed.

5 Conclusion

In this paper, wavelet transform is introduced into the field of multimodal intent recognition for the first time, and a novel wavelet-based multimodal intent recognition method, WDMIR, is proposed. The method serializes multilevel decomposition of video and audio data by wavelet transform and realizes the extraction and fusion of video and audio features in the frequency domain. Through the collaborative fusion module, it realizes the collaborative representation from bimodal to trimodal, which improves the expression of deep features and better identifies the semantic information in multimodal data. Through the progressive fusion module, the whole different levels of feature representations are effectively represented to improve the accuracy of multimodal intent recognition. Comprehensive experimental results show that the proposed wavelet-driven approach can enhance the performance of the model in multimodal intention recognition tasks. This study not only demonstrates the potential of wavelets in multimodal learning but also provides a new perspective on feature extraction and fusion of multimodal data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 62337001, U23A20319, 62406303), the Key Technologies R&D Program of Anhui Province (No. 202423k09020039), the Anhui Provincial Natural Science Foundation (No. 2308085QF229), and the Fundamental Research Funds for the Central Universities. Additional support was provided by the Youth Science and Technology Talent Growth Project of Guizhou Provincial Department of Education (Grant No. KY[2022]054), the Guizhou Provincial Higher Education Undergraduate Teaching Content and Curriculum System Reform Project (Grant No. GZJG2024331), and the Guizhou Provincial Science and Technology Projects (Grant No. QKHJC[2024]Youth012).

References

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[Chen *et al.*, 2021] Luefeng Chen, Zhentao Liu, Min Wu, Kaoru Hirota, and Witold Pedrycz. Multimodal emotion recognition and intention understanding in human-robot interaction. *Developments in Advanced Control and Intelligent Automation for Complex Systems*, pages 255–288, 2021.

[Frusque and Fink, 2024] Gaetan Frusque and Olga Fink. Robust time series denoising with learnable wavelet packet transform. *Advanced Engineering Informatics*, 62:102669, 2024.

[Graves *et al.*, 2006] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020.

[Huang *et al.*, 2023] Xuejian Huang, Tinghuai Ma, Li Jia, Yuanjian Zhang, Huan Rong, and Najla Alnabhan. An effective multimodal representation and fusion method for multimodal intent recognition. *Neurocomputing*, 548:126373, 2023.

[Huang *et al.*, 2024] Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10206–10210. IEEE, 2024.

[Lahmiri, 2014] Salim Lahmiri. Wavelet low-and high-frequency components as features for predicting stock prices with backpropagation neural networks. *Journal of King Saud University-Computer and Information Sciences*, 26(2):218–227, 2014.

[Lee and Toutanova, 2018] JDMCK Lee and K Toutanova. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 3(8), 2018.

[Li *et al.*, 2023] Guanxin Li, Jingang Shi, Yuan Zong, Fei Wang, Tian Wang, and Yihong Gong. Learning attention from attention: Efficient self-refinement transformer for face super-resolution. In *IJCAI*, pages 1035–1043, 2023.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2023] Ye Liu, Kai Zhang, Zhenya Huang, Kehang Wang, Yanghai Zhang, Qi Liu, and Enhong Chen. Enhancing hierarchical text classification through knowledge graph integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5797–5810, 2023.

[Loshchilov, 2017] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Phutke *et al.*, 2023] Shruti S Phutke, Ashutosh Kulkarni, Santosh Kumar Vipparthi, and Subrahmanyam Murala. Blind image inpainting via omni-dimensional gated attention and wavelet queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1260, 2023.

[Qiu, 2024] Lingling Qiu. Application of deep learning-based user intent recognition in human-computer interaction. In *2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, pages 549–552. IEEE, 2024.

[Rahman *et al.*, 2020] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.

[Sabry *et al.*, 2024] Ahmad H Sabry, Omar I Dallal Bashi, NH Nik Ali, and Yasir Mahmood Al Kubaisi. Lung disease recognition methods using audio-based analysis with machine learning. *Heliyon*, 2024.

[Saha *et al.*, 2020] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, 2020.

[Satirapod *et al.*, 2001] Chalermchon Satirapod, Clement Ogaja, Jinling Wang, and Chris Rizos. An approach to gps analysis incorporating wavelet decomposition. *Artificial Satellites*, 36(2):27–35, 2001.

[Soleymani *et al.*, 2017] Mohammad Soleymani, Michael Riegler, and Pål Halvorsen. Multimodal analysis of image search intent: Intent recognition in image search from user behavior and visual content. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 251–259, 2017.

[Sun *et al.*, 2024] Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26963–26973, 2024.

[Sun *et al.*, 2025] Xinjie Sun, Kai Zhang, Qi Liu, Shuanghong Shen, Fei Wang, Yuxiang Guo, and Enhong Chen. Daskt: A dynamic affect simulation method for knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

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

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[Zhang *et al.*, 2019] Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5773–5780, 2019.

[Zhang *et al.*, 2021] Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):377–389, 2021.

[Zhang *et al.*, 2022a] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697, 2022.

[Zhang *et al.*, 2022b] Kai Zhang, Qi Liu, Zhenya Huang, Mingyue Cheng, Kun Zhang, Mengdi Zhang, Wei Wu, and Enhong Chen. Graph adaptive semantic transfer for cross-domain sentiment classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1566–1576, 2022.

[Zhang *et al.*, 2022c] Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2203.16369*, 2022.

[Zhang *et al.*, 2024] Yanghai Zhang, Ye Liu, Shiwei Wu, Kai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. Leveraging entity information for cross-modality correlation learning: The entity-guided multimodal summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9851–9862, 2024.

[Zhou *et al.*, 2024] Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17114–17122, 2024.

[Zou *et al.*, 2022] Yicheng Zou, Hongwei Liu, Tao Gui, Junzhe Wang, Qi Zhang, Meng Tang, Haixiang Li, and Daniel Wang. Divide and conquer: Text semantic matching with disentangled keywords and intents. *arXiv preprint arXiv:2203.02898*, 2022.