information

MDPI

Article

Attention Bidirectional Gated Fusion Based Multimodal Intent Recognition Under Uncertain Missing Modalities

Ling Shang [1], Zhizhong Liu [2,3,*], Yuxuan Wu [2], Xiaoyu Song [4], Jian Yu [5] and Quan Z. Sheng [6]

[1] Digital Smart Creative Department, Henan Culture and Tourism Investment Group Co., Ltd., Luoyang 471026, China; shangl@hnwltzjt.cn
[2] School of Computer and Control Engineering, Yantai University, Yantai 264005, China; wuyuxuanpc@126.com
[3] Shandong Key Laboratory of Digital Service Computing and Systems, Yantai 264005, China
[4] College of Information Engineering, Yantai Institute of Science and Technology, Yantai 264005, China; sxyytdxjkxy@gmail.com
[5] Department of Computer Science, Auckland University of Technology, Auckland 1142, New Zealand; jian.yu@aut.ac.nz
[6] School of Computing, Macquarie University, Sydney, NSW 2109, Australia; michael.sheng@mq.edu.au
* Correspondence: zhizhongliu@ytu.edu.cn

Abstract

Currently, uncertain missing modalities pose new challenges to multimodal intent recognition. To tackle this issue, this work proposes an Attention Bidirectional Gated Fusion Based Multimodal Intent Recognition model under Uncertain Missing Modalities (named ABGFMIR). Firstly, ABGFMIR extracts the features of each modality (text, audio, visual) with the LSTM network, respectively. Secondly, ABGFMIR narrows the distances between audio, visual and text modality based on Central Moment Discrepancy (CMD), and then performs multimodal feature fusion through an attention bidirectional gated fusion method. Then, corresponding attention level prompts are generated based on the uncertain missing modalities situations of the current sample. The fused multimodal features are then input into the Transformer encoder and decoder, and the prompts are injected into the keys and values of the multihead self-attention layer to guide the Transformer to focus on the missing modes and dynamically adjust the attention distribution, enhancing the robustness of ABGFMIR to different missing modes. Finally, the features produced by the Transformer are fed into the classification layer for intent recognition. Simultaneously, the pre-trained model (AGNN) that trained with the complete modality is employed in the classification layer to guide the main module of ABGFMIR. Two public benchmark datasets (MIntRec and EMOTyDA) are adopted for performance verification. Compared with the other five baseline models, on the MIntRec dataset, ABGFMIR improved accuracy by an average of 2.68 and improved F1 values by an average of 3.24. On the EMOTyDA dataset, ABGFMIR improved accuracy by an average of 2.28 and improved F1 values by an average of 3.44.

Keywords: multimodal intent recognition; uncertain missing modalities; transformer; LSTM; multimodal feature fusion

Academic Editors: Zoran H. Perić, Tomas Skovranek, Vlado Delić, Vladimir Despotovic and Stefan Panic

Received: 23 May 2026
Revised: 24 July 2026
Accepted: 24 July 2026
Published: 28 July 2026

1. Introduction

Intent recognition is an important technology that is dedicated to determining users' intents through learning user-related data. Intent recognition techniques can enable systems to understand user's needs; thus providing services timely and accurately [1], which has important application in multiple fields, such as dialogue systems, chatbots, intelligent

customer assistants, and so on. Initially, some research has been conducted on intent recognition only based on text data, and has obtained some wonderful results [2–4]. Recently, with the emergence of multimodal data (including text, audio and visual) and their advantage of information complementarity, multimodal intent recognition (MIR) has become an increasingly hot topic and has attracted considerable attention [5].

Multimodal intent recognition aims to identify users' intents accurately through learning the multimodal data simultaneously [6]. In the past few years, some effective methods for MIR have been proposed, such as multimodal intent recognition with saliency and text-guided fusion [7], multimodal intent recognition based on mamba-evidence-driven triplet contrastive learning [8], and multimodal intent recognition based on text-guided frequency-decoupled modeling [9]. Moreover, ref. [10] developed a modal mapping coupling and gate-driven contrastive learning approach for multimodal intent recognition. Sun et al. [11] proposed a context-augmented global contrast (CAGC) method to capture rich global context features by mining both intra-and cross-video context interactions for MIR.

However, existing research on MIR always assumes that the three modalities (text, audio, visual) are available all the time. In actuality, in real-world environments, it is not possible to obtain the three modalities all the time, due to some uncontrollable reasons. Therefore, uncertain missing modalities frequently occur in MIR, which may include situations where no modality is missing, any one modality is missing, or any two modalities are missing. For example, as shown in Figure 1, when the camera is broken, the visual modality will be missing. Moreover, speech and text will be missing due to privacy protection policies or the recording device being broken.

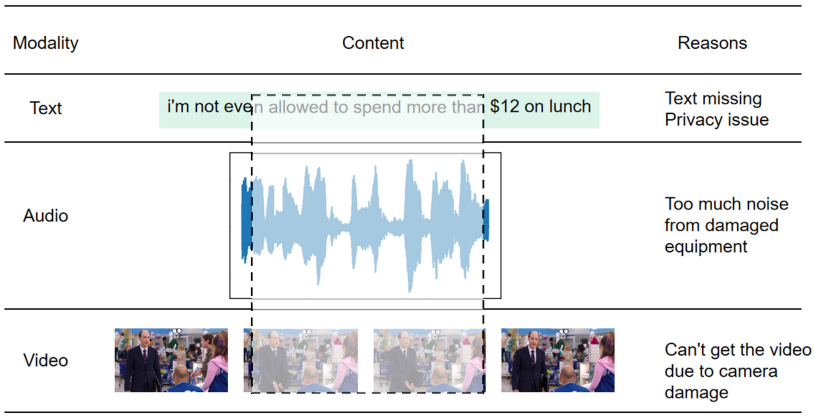

Figure 1. An example of missing modalities.

For MIR under uncertain missing modalities, models trained with the full modalities will become invalid. Therefore, determining how to realize MIR uncertain missing modalities becomes a new challenge. However, to the best of our knowledge, currently few research works have considered MIR under uncertain missing modalities. Although some models have been proposed for multimodal sentiment analysis under uncertain missing modalities [12–15], these models cannot be directly applied for MIR under uncertain missing modalities. In all, there are some critical challenges to be addressed for MIR under uncertain missing modalities, which are presented as follows:

- **How can we enhance the quality of multimodal fusion for MIR under uncertain missing modalities?** Currently, some research work has demonstrated that the text modality plays a major role in MIR. However, existing works for MIR treat all the three modalities equally in fusion, and cannot deeply interact with multimodal features in the fusion, thus affecting the quality of multimodal fusion.

- **How can we overcome the semantic inconsistency caused by uncertain missing modalities for MIR?** Uncertain missing modalities cause the semantics of multimodal samples to be inconsistent with the semantics of complete modalities. Existing models trained with full modalities cannot overcome the semantic inconsistency. Therefore, determining how to handle the problem of semantic inconsistency caused by uncertain missing modalities becomes a challenge for MIR.
- **How can we handle the uncertain missing modalities for MIR?** Existing works on MIR usually assume that all the modalities are available all the time, failed to consider the uncertain missing modalities. Therefore, determining how to effectively handle the uncertain missing modalities for MIR becomes a new challenge.

To address the above issues, this work proposes an Attention Bidirectional Gated Fusion Based Multimodal Intent Recognition under Uncertain Missing Modalities (named ABGFMIR). ABGFMIR first extracts features of each modality with LSTM. Then, the extracted features are fused with the attention-based bidirectional gated features fusion module, where the distance constraint strategy based on CMD is used to bring each modality closer to the text modality. Next, the fused multimodal features are fed into the Transformer for learning. Meanwhile, different attention level prompts are input according to the uncertain modal absence, thus directing the network to pay attention to those missing modalities. Finally, the output of the Transformer is used for final intent recognition. At the same time, the pre-trained model trained with the full modalities is used to guide the model at the classification layer during the training process, thus helping the model to avoid overfitting and improve its generalization ability. The main contributions of this work are summarized as follows:

- **To enhance the quality of multimodal fusion**, we propose an attention-based bidirectional gated multimodal feature fusion module. Moreover, inspired by the fact that text modality plays a dominant role in MIR, we propose to reduce the distance from other modalities to the text modality to improve the quality of multimodal fusion.
- **To address the problem of semantic inconsistency caused by uncertain missing modalities**, we propose learnable Transformer prompts to mitigate the performance degradation caused by missing modalities. In this work, different types of attention level prompts are input according to different scenarios of missing modalities, thus enabling the model toward the target domain through training with fewer parameters, guiding the model to focus on the missing modalities.
- **To better cope with uncertain missing modalities**, we introduce a pre-trained model (Attention-based Bidirectional Gated Neural Network, AGNN) that is trained with complete modalities and is used to provide effective guidance for the final classification in scenarios of uncertain missing modalities, thereby improving the robustness of ABGFMIR.

Our work is organized as follows: Section 2 reviews related works. Section 3 introduces our proposed model. Section 4 presents experimental results and analysis. Section 5 summarizes our work.

2. Related Work

Intent recognition based on text modal data is a crucial task in spoken language understanding [16]. Considering the close correlation between intent recognition and slot filling, most works adopt a federated model to exploit shared knowledge across tasks. Qin et al. [17] proposed a collaborative interactive transformer model to model two-way connections that simultaneously perform the two tasks of intent recognition and slot filling in a unified framework. Wu et al. [18] proposed a new nonautoregressive model

SlotRefine for joint intent detection and slot filling. In real-life scenarios, users often have multiple intents in the same utterance. To address this problem, Qin et al. [19] proposed an adaptive graph interaction framework for multi-intent detection, introducing an intent-slot graph interaction layer that can accurately extract relevant intent information and simulate the strong association between slots and intents. Significant advances in intent recognition rely on large amounts of labeled training data, which cannot work in low-resource environments, so several works have studied intent recognition in low-resource situations. In terms of few-shot learning, Hou et al. [20] explored the few-shot multilabel problem in intent recognition and proposed a meta-calibration threshold mechanism with kernel regression and logic adaptation, which uses previous domain experience, and new domain knowledge is used to estimate the threshold. In terms of zero-shot learning, Wu et al. [21] proposed a label-aware BERT attention network (LABAN) based on BERT for zero-shot multi-intent recognition.

In terms of computer vision, Jia et al. [22] proposed an object/context localization loss that focuses the model on key areas of the image to identify the intent behind social media images. Joo et al. [23] proposed a hierarchical model based on the syntactic attribute layer to identify communicative intents in politicians' photos. Fang et al. [24] identified pedestrians' and cyclists' lane-changing intents based on their arm movements in images. Yang et al. [25] effectively solved the problem of identifying pedestrians' intent to cross the road in an urban environment by introducing a search target module, an action recognition module, and a distance encoding module. Xu et al. [26] proposed the IHOI framework model, which uses visual information to identify human intent in human–computer interaction.

In terms of multimodal intent recognition, Zhang et al. [27] showed, through a large number of experiments, that compared with pure text modality, the performance of the model using multimodality was significantly improved, proving the effectiveness of using multimodal information for intent recognition. Maharana et al. [28] proposed a multimodal cross-attention model that combines video modality and text modality for intent recognition. In terms of social network intent recognition, Kruk et al. [29] combined text and pictures to identify the intent of people posting on Instagram, but modal fusion is just a simple projection addition, without considering the differences and complementarity between different modalities. Kiela et al. [30] studied hateful intent in social network posts by adding text to images to construct benign confounders, and found that state-of-the-art methods performed poorly compared to humans. Liu et al. [31] used a multihead cross-attention mechanism and a graph neural network to establish a hierarchical framework to identify the satirical intent of social network posts from the perspective of multigranularity alignment.

Aiming at the problem of identifying the intent of self-media advertising on social networks, Zhang et al. [32] proposed a new supervised neural autoregressive model, which can effectively learn hidden features and extract features through a graph convolution network to improve the recognition accuracy. Huang et al. [12] proposed an effective multimodal representation and fusion method for intent recognition for complex real-world multimodal scenes. By combining text, audio and visual features, using modality-specific and shared encoders, and an adaptive multimodal fusion method, the method outperforms state-of-the-art methods.

3. Methodology

To solve the problem of multimodal intent recognition (MIR) under uncertain missing modalities, this work proposes an Attention Bidirectional Gated Fusion Based Multimodal Intent Recognition model considering Uncertain Missing Modalities (named as ABGFMIR). In this section, we first introduce the definition of the problem studied in this work.

Then, we describe the workflow of ABGFMIR. Finally, the key modules of ABGFMIR are presented in detail.

3.1. Problem Statement

Given multimodal data $D = \{x_t, x_a, x_v\}$ including three modalities, where x_t, x_a and x_v denote the text, audio, and visual modality, in this work, we use x_m^l to represent the uncertain missing modality, where $m \in v, a, t$. For example, when the audio modality is missing, the multimodal data can be presented as $\left[x_t, x_a^l, x_v\right]$. In our study, we investigate how to perform MIR under uncertain missing modalities (that is, conducting intention recognition based on data $D = \left\{x_t, x_a, x_m^l\right\}$), where uncertain missing modalities mean that the number and types of missing modalities are uncertain.

3.2. Model Overview

The structure of our proposed model ABGFMIR is illustrated in Figure 2, which consists of two modules, which are the pre-trained AGNN model (Attention-based Bidirectional Gated Neural Network) that is trained with the full modalities, and the main AGNN model that will be trained with uncertain missing modalities. In this work, to better cope with uncertain missing modalities, the pre-trained AGNN can transfer the learned knowledge to the classification layer, providing effective guidance for the final classification in scenarios of missing modalities, thereby improving the robustness of ABGFMIR in different modalities missing scenarios. In the following sections, we will introduce ABGFMIR with the hypothesis that the audio modality is missing.

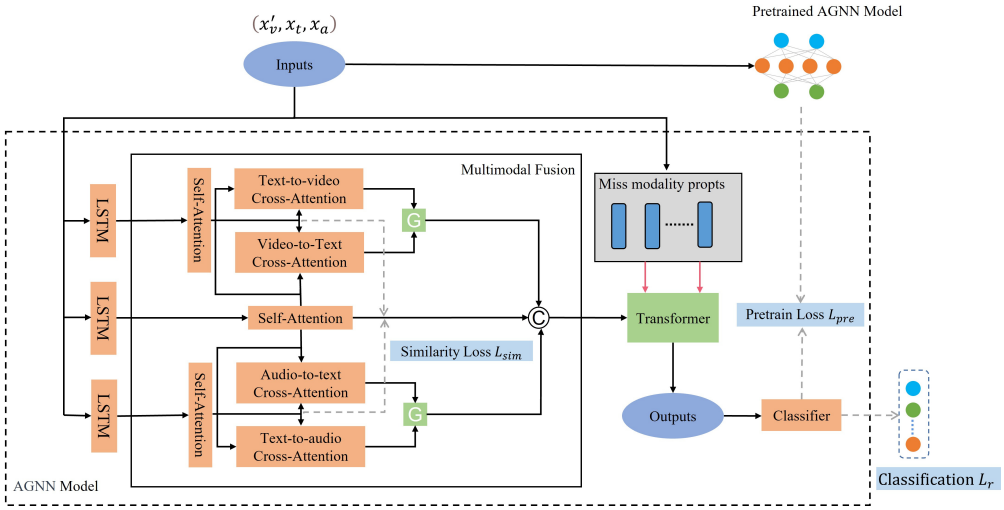

Figure 2. Structure of the ABGFMIR model.

The workflow of ABGFMIR can be described as follows. Firstly, the multimodal data $\left[x_t, x_a^l, x_v\right]$ with uncertain missing modalities is input into the pre-trained AGNN model, and obtains the embedding $\left[x_t, x_a^l, x_v\right]$ produced by the pre-trained AGNN model. The structure of the pre-trained AGNN model is the same as the main AGNN model. Meanwhile, the multimodal data $\left[x_t, x_a^l, x_v\right]$ with uncertain missing modality is also input into the main AGNN model. In this model, the data of each modality is input into an LSTM network to extract features of each modality. Then, features of each modality are fused in the fusion module to obtain the joint features of the three modalities. Then, similarity learning is performed to narrow the distance between the audio, video and the text. Next, the Transformer encoder module learns the joint feature representation and adds different modal missing prompt blocks before the original key K and value V of the Transformer

according to the missing modality scenarios. Finally, the output of the Transformer is used to classify the intent with the classifier. At the same time, the output of the pre-trained model is used to perform supervised learning on the entire model at the classification layer, thereby improving the generalization ability of ABGFMIR. In the following sections, we will introduce the main modules of ABGFMIR.

3.3. Extracting Features of Each Modality with LSTM

In actuality, the three modalities (text, audio, visual) used for intent recognition are temporal data. Moreover, LSTM is suitable for processing time-series data, and can capture the temporal information in text, audio and visual. In this work, we adopt LSTM to extract the features of each modality. Specifically, the audio modality and visual modality are processed by LSTM; thus, we obtain features E_a and E_v from the original data x_a and x_v. The text modality is also encoded with LSTM, and we obtain the discourse level text features E_t from the original data x_t. The process of features extraction for each modality is described using Equation (1):

$$E_m = LSTM\left(x_m, \theta_m^{lstm}\right) \tag{1}$$

where $x_m (m \in t, v, a)$ denotes each modality, and $\theta_m^{lstm} (m \in t, v, a)$ denotes the parameters of the LSTM model.

3.4. Attention-Based Bidirectional Gated Fusion

ABGFMIR performs multimodal feature fusion after extracting features of each modality, and the fusion of multimodal features can provide more information for MIR; thus, it can improve the accuracy of intention recognition. However, existing works regard the three modalities as equal when fusing multimodal features, and cannot interact the three modalities deeply. To solve this problem, we propose a multimodal feature fusion module based on the attention-based bidirectional gated neural network to achieve deep fusion of the three modalities. The structure of this module is shown in Figure 3, which consists of self-attention, similarity learning, cross-modal attention and gated bidirectional fusion.

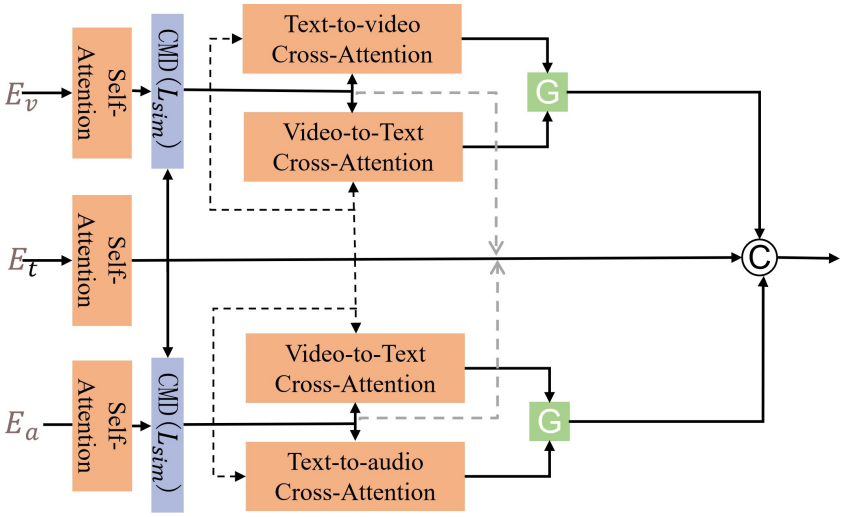

Figure 3. Multimodal feature fusion module.

In the multimodal feature fusion module, the self-attention mechanism is first executed on the features of each modality. The self-attention mechanism can compute the global information of the input sequence without being limited to local regions. Moreover, the self-

attention mechanism is used separately for each modality, and this can avoid information confusion. The above calculation can be described as Equation (2):

$$H_m = SelfAttention(E_m) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

where E_m means features of each modality, $m \in t, v, a$. $Q = W^Q E_m$, $K = W^K E_m$, and $V = W^V E_m$ denote the Query, Key, and Value matrices, respectively. W^Q, W^K and W^V represent the parameters that will be learned during the training process. $\sqrt{d_k}$ is the dimension of the key matrix.

To enhance the robustness of the joint multimodal features, in this work, we propose a method for inter-modal feature similarity learning, aiming to mitigate the modal differences during the prediction of missing modalities. As text modality plays a major role in multimodal intent recognition [7,9], we propose similarity learning for text–audio and text–video. The quality of multimodal feature fusion is improved by minimizing the similarity loss to make the video modality and audio modality more close to the text modality.

In this work, we use the Central Moment Discrepancy (CMD) distance as the similarity loss. The CMD distance evaluates the dissimilarity between two distributions by comparing the distance of difference between their central moments. Let X and Y be random samples with probability distributions p and q on the interval $[a, b]$. The CMD distance metric can be calculated using Equation (3):

$$CMD_K(X, Y) = \frac{1}{|b-a|} \parallel E(x) - E(Y) \parallel_2 + \sum_{k=2}^{k} \frac{1}{|b-a|^k} \parallel C_k(X) - C_k(Y) \parallel_2 \tag{3}$$

where $E(x) = (1/|X|)\sum_{x \in X} x$ is the empirical expectation vector of the sample x, and $C_k(x) = E\left((x - E(x))^k\right)$ is a vector of all kth-order sample central moments of the x-coordinate as shown in Equation (4).

$$L_{sim} = \frac{1}{2} \sum_{(h_1, h_2) \in (t,a),(t,v)} CMD_K(h_1, h_2) \tag{4}$$

where h_1, h_2 are features calculated with self-attention, $(h_1, h_2) \in (t, a), (t, v)$.

Next, cross-modal attention is used to learn correlation features between different modalities. The main difference between cross-modal attention and self-attention is the different orders of inputs for cross-attention. Considering that the audio modality and video modality are not as important as the text modality, cross-modal features of audio-to-video and video-to-audio are not computed. In this work, cross-modal attention is used to learn the text-to-vision related features, vision-to-text related features, text-to-audio related features, and audio-to-text related features, which can be calculated using Equation (5).

$$C_{h_1 \to h_2} = softmax\left(\frac{Q_{h_2} K_{h_1}^\top}{\sqrt{d_k}}\right)V_{h_1} \tag{5}$$

where $(h_1, h_2) \in \{(t, a), (t, v), (a, t), (v, t)\}$, $Q_{h_2} = W^Q_{h_1 \to h_2} h_2$, $K_{h_1} = W^K_{h_1 \to h_2} h_1$, and $V_{h_1} = W^V_{h_1 \to h_2} h_1$. $W^Q_{h_1 \to h_2}$, $W^K_{h_1 \to h_2}$, and $W^V_{h_1 \to h_2}$ denote the learnable projection parameters of the respective cross-modal attention learning.

After that, audio-related features $C_{h_t \to h_a}$ and $C_{h_a \to h_t}$ are fused with the bidirectional gate. Meanwhile, video-related features $C_{h_v \to h_t}$ and $C_{h_t \to h_v}$ are fused. Bidirectional gated fusion (as shown in Figure 4) has an advantage in processing temporal data, especially for tasks that require consideration of contextual relationships. The use of this mechanism can

improve the expressive power of the model and capture the key information in the sequence data more effectively. The final features of the audio and video modalities are obtained using bidirectional gated fusion, which can be described using Equations (6) and (7):

$$H_a^t = G_{c_a,c_t}^U \otimes C_{a\to t} + C_{t\to a} \tag{6}$$

$$H_v^t = G_{c_v,c_t}^U \otimes C_{v\to t} + C_{t\to v} \tag{7}$$

where $\otimes$ refers to the tensor product, the G^U gate function is used to control the information flow, and H_a^t and H_v^t represent the audio features and video features computed by the hidden layer. Taking the acquisition of audio feature H_a^t as an example, the gate function operation is defined as Equation (8):

$$G_U^{i,j} = Sig(Conv(cat(i,j))) \tag{8}$$

where *cat* denotes the splicing operation, *Conv* denotes the convolutional layer and *Sig* denotes the Sigmoid function of the tensor. The gate function first splices the two input parameters, followed by a convolution operation on the spliced result; finally, the final output is obtained with the Sigmoid function.

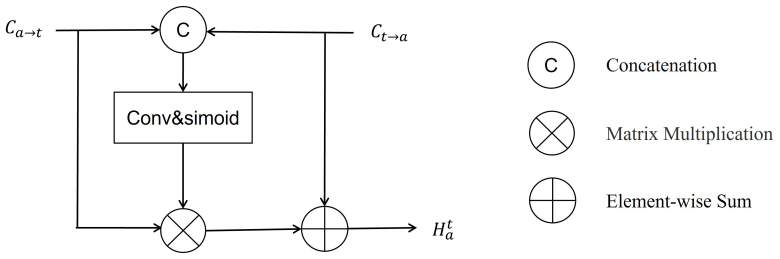

Figure 4. Bidirectional gating.

Finally, we concatenate the feature vectors of all modalities to obtain a common joint representation of the three modalities H_{all}, which can be calculated according to Equation (9):

$$H_{all} = \left[H_t \parallel H_a^t \parallel H_v^t\right] \tag{9}$$

where $\parallel$ denotes the concatenation operation. The concatenated features are further projected into a shared d-dimensional space through a linear projection layer, yielding $H_{all} \in R^{l\times d}$. H_t is the text feature extracted by self-attention described in Equation (2), H_a^t is the result of gated fusion of audio features calculated with Equation (6), and H_v^t is the result of gated fusion of video features calculated with Equation (7).

3.5. Modality Missing Prompt Based on Transformer

To handle the uncertain missing modalities in MIR, it is necessary to make ABGFMIR aware of missing modalities during the training process. Inspired by the works [33,34], we propose a modality missing prompt module-based Transformer. The main idea of this module is to add different trainable vectors to the keys and values of the multihead self-attention layer of each block in the Transformer according to missing modality situations, so as to enhance the performance of the ABGFMIR model for handling missing modalities.

In the multihead self-attention layer, the original Key, Value and Query are computed as $K = H_{all}$, $V = H_{all}$, $Q = H_{all}$, respectively. $H_{all} \in R^{l\times d}$ is the input sequence of the training samples, l is the length, and d denotes the hidden dimension of Transformer. To learn the specific information of different modal missing cases, there are 7 trainable d-dimensional vectors as key and value supercues, denoted as $P_m \in R^{h\times d}$, h is the length of the sequence, d is the model dimensionality, and m is the case of missing modality,

which denotes the complete modal inputs and the 6 different modal missing prompts, respectively, which can be illustrated in Figure 5. The calculation process is described as Equations (10) and (11):

$$K' = \text{Concat}(P_m, H_{\text{all}}) \tag{10}$$

$$V' = \text{Concat}(P_m, H_{\text{all}}) \tag{11}$$

where $Concat(\cdot)$ operates along the feature dimension, new $V'(K') \in R^{(h+l)\times d}$ is used to compute the multihead self-attention. Next, K', V' and Q with missing modality prompts on the splice are ready to be fed into the Transformer for computation with the multihead attention mechanism, which is illustrated in Figure 5.

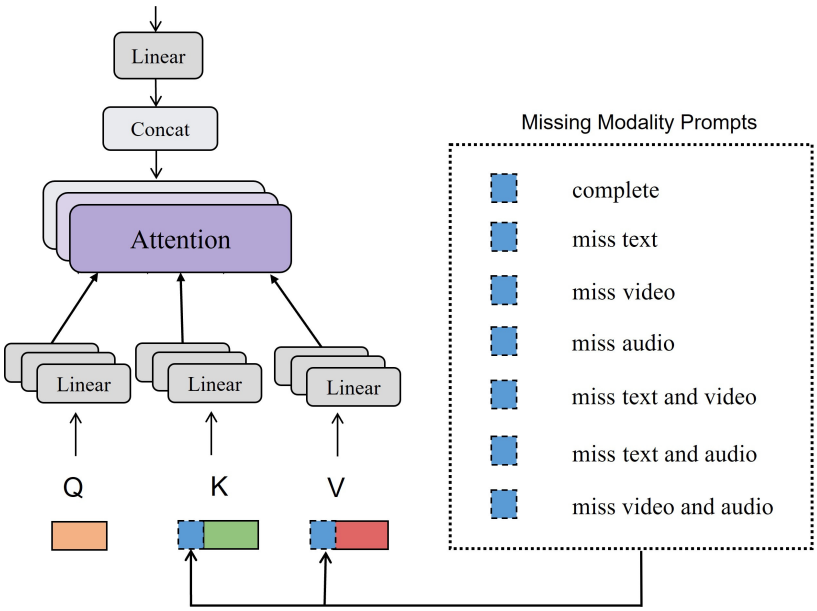

Figure 5. Modality missing prompts.

Next, the features obtained from the multimodal feature fusion module are input into the Transformer module, and the modality missing prompts are injected into K and V of the multihead self-attention layer according to the modality missing situations. The computation process can be described using Equations (12)–(14):

$$H_n = MultiHead(Q, K', V') = MultiHead(H_{all}, [H_{all}||P_m], [H_{all}||P_m]) \tag{12}$$

$$H_n = Layernorm(H_{all} + H_n) \tag{13}$$

$$H_n = Relu\left(H_n W_e^1 + b_e^1\right) W_e^2 + b_e^2 \tag{14}$$

where w_1 and w_2 are weight matrices, b_1 and b_2 are biases, and $ReLU$ denotes the activation function. Similarly, using the output H_n of the encoder as the input to the decoder, the representation of the decoded output H_d can be expressed using Equations (15)–(17):

$$H_d = MultiHead(H_n, H_n, H_n) \tag{15}$$

$$H_d = Layernorm(H_d + H_n) \tag{16}$$

$$H_d = Relu\left(H_d W_d^1 + b_d^1\right) W_d^2 + b_d^2 \tag{17}$$

where w_1 and w_2 are weight matrices, b_1 and b_2 are biases, and $ReLU$ denotes the activation function.

3.6. Intent Recognition

To realize the intent recognition, the output feature H_d is fed into the classifier to obtain the intent category. For the classifier, the classification loss is defined using Equations (18) and (19):

$$\hat{y} = W_1 ReLU(W_2 H_d + b_2) + b_1 \tag{18}$$

$$L_r = CrossEntropyLoss(\hat{y}, y) \tag{19}$$

where CrossEntropyLoss denotes the loss function, W_1 and W_2 are the weight matrices, b_1 and b_2 are the biases, and ReLU denotes the activation function.

3.7. Generalization Learning Based on Pre-Trained Model

For handling the uncertain missing modalities, inspired by existing work [35], we propose a supervised module for generalizability learning based on a pre-trained model. The goal of this module is to allow the model to learn the generalization ability of a pre-trained model which is trained with full modalities, and the results obtained will be better than those of the model trained with data under uncertain missing modalities. Specifically, ABGFMIR introduces the distance between the pre-trained model and the main AGNN model through the pre-training loss in the classification layer. The output of the pre-trained model contains more information compared to the labels, and passes the semantic knowledge of the complete modality to the model when the uncertain modality occurs, so, it can effectively improve the generalization ability of ABGFMIR. Moreover, the pre-trained AGNN model and the main AGNN model are two independent models which have the same structure and do not share parameters. The loss function for pre-training is defined using Equation (20):

$$L_d = D_{\mathrm{KL}}\left(\hat{y}_{pre}/\tau \parallel \hat{y}/\tau\right) \tag{20}$$

where D_{KL} denotes the KL dispersion and τ denotes the temperature coefficient, which measures the difference between two probability distributions. $\hat{y}_{pre}$ is the output of the pre-trained model and $\hat{y}$ is the main AGNN model output.

3.8. Model Training

Our proposed model ABGFMIR consists of several computation parts, including intent recognition, similarity learning, and pre-training guidance. To train ABGFMIR fully, each computation part has a corresponding loss function. ABGFMIR model training requires combining the loss functions of all the tasks to be trained together. The overall training loss function is defined as Equation (21):

$$L_{total} = \lambda_1 L_r + \lambda_2 L_{sim} + \lambda_3 L_{pre} \tag{21}$$

where λ_1, λ_2, λ_3 are the corresponding weights, L_r is the classification loss, L_{sim} is the similarity loss, and L_{pre} is the pre-training loss.

4. Experiments

To validate the performance of the ABGFMIR model, we conduct extensive experiments based on two public datasets MIntRec [27] and EMOTyDA [36]. In this section, we first introduce the two public benchmark datasets, then we introduce the experimental environment and parameter settings. Next, we introduce the five baseline models. Finally, we analyze the results of comparison experiments, ablation experiments, and other related experiments.

4.1. Benchmark Datasets

In this work, two public benchmark datasets, MIntRec [27] and EMOTyDA [36], are adopted. Both datasets are multimodal datasets including visual, text and audio modalities. For the MIntRec dataset, the original data in this dataset is derived from the TV series Superstore. MIntRec is currently the first benchmark dataset for real-world multimodal intent recognition and contains 2224 instances of text, video, and audio modalities. MIntRec contains 2 coarse-grained classifications and 20 fine-grained intent classifications, and the data is divided according to 3:1:1 into training, a validation and a test set. In contrast to MIntRec, EMOTyDA is a large-scale dataset for conversation act classification, where all data are labeled with 12 common conversation act labels. Due to the limitation of computational resources, in this work, 2500 representative samples are selected as the experiment dataset on the basis of EMOTyDA dataset, which is divided into training set, validation set and test set with the ratio of 3:1:1. To ensure fairness, all the models used for performance comparison are verified on these two benchmark datasets.

In this work, to generate uncertain missing modalities, we conduct random missing operation for each modality of each sample in the training, validation, and testing sets. The specific operation is to generate a binary mask matrix for each modality in advance, where 0 represents that the modality is missing and 1 represents that the modality is fully preserved. These masks are sampled and output according to the preset missing probability, ensuring that the overall missing rate accurately matches the set value. This work uniformly sets the missing rates of the training set, validation set, and test set to 0%, 10%, 20%, 30%, 40%, and 50% to ensure the consistency and reproducibility of the experiment. To our knowledge, research works typically use the above-mentioned methods to generate uncertain missing multimodal data for performance verification.

4.2. Experimental Environment and Parameter Settings

All the experiments were conducted on a Linux server with operating system version Ubuntu 18.04, Intel(R) Xeon(R) Platinum 8255C with 2.50 GHz CPU, one GeForce RTX 3090 GPU and 128 GB memory. The programming language is Python (version 3.8). The deep learning architecture is PyTorch 1.11.0. The experiments use a CUDA-based driver for GPU parallel computing and accelerated deep learning with Cuda version 11.3. A summary of parameters is shown in Table 1.

Table 1. Detailed parameter settings in all experiments.

Hyperparameter	Symbol	Value
Batch size	b	16
Epoch size	e	100
Discard rate	p	0.1
Hidden layer size	d	256
Uncertainty modal missing rate	η	[0–50%]
Learning rate	l_r	0.0003
Maximum text length	n_t	30
Maximum audio length	n_a	230
Maximum video length	n_v	480
Training weights	$\lambda_1, \lambda_2, \lambda_3$	0.8, 0.5, 0.5

4.3. Baseline Models

To validate the performance of our proposed model ABGFMIR, we selected five state-of-the-art multimodal intent recognition models for comparison, which are introduced as follows:

Mult [37]: MulT uses an end-to-end approach to process unaligned multimodal data. It extends the normal Transformer to a cross-modal Transformer through a bidirectional cross-modal attention mechanism, which helps to capture interactions between different modalities in the potential space.

MAG_Bert [38]: MAG_Bert model designs a multimodal adaptive gate MAG that allows the BERT to accept multimodal nonverbal data during fine-tuning. The MAG can generate positional shifts in semantic space to accommodate audio and visual information. The MAG_Bert can be flexibly configured between the BERT layers to receive inputs from nonverbal modalities.

MISA [39]: MISA is able to learn complementary information to minimize redundancy and incorporate diverse information sets. MISA adopts a multitask learning framework, which learns different modalities by projecting them into two spaces, the modality-invariant subspace and the modality-specific subspace, and, thus, obtains multimodal representations with both modality-invariant and modality-specific attributes. On the one hand, MISA learns the features common to all modalities by using the shared subspace; on the other hand, different subspaces are designed to capture the unique properties of each modality. To this end, the training objective consists of four aspects: similarity loss, difference loss, reconstruction loss and task-specific loss.

IF-MMIN [40]: The IF-MMIN approach addresses the challenges posed by differences between modalities by learning invariant features and using them to predict and fill in missing modal data, especially when modal data is incomplete or uncertain. In a large number of experiments, the model performs well and outperforms the baseline approach, significantly improving multimodal emotion recognition performance.

EMRFM [12]: EMRFM is an effective multimodal representation and fusion method for intent recognition in real multimodal scenes. EMRFM constructs modal-shared and modal-specific encoders to learn modal-shared and modal-specific feature representations, and designs an adaptive multimodal feature fusion method based on attentional gated neural networks to eliminate noisy features. Experiments show that the multimodal representation of the EMRFM model learns the shared and specific features of modalities well, and the multimodal feature fusion of the model realizes adaptive fusion, which effectively reduces the possible noise interference.

4.4. Hyperparametric Sensitivity Analysis

To find the appropriate hyperparameters for our proposed model, we conduct the hyperparameter sensitivity analysis experiments. In this experiment, F1 score is used as an evaluation index; the multitask learning parameters of model training λ_1, λ_2, λ_3 and the temperature coefficient τ of the pre-training module are experimented upon. In the hyperparameter sensitivity analysis experiments, because there are too many hyperparameters, we first analyze the hyperparameters of multitask learning λ_1, λ_2, λ_3, and then analyze the temperature coefficients of the pre-training module in the case of the fixed multitask learning parameters λ_1, λ_2, λ_3. coefficients τ.

For the multitask learning parameters λ_1, λ_2, λ_3, we do not adjust the three hyperparameters of multitask learning. In contrast, this paper first randomly fixes two of the three hyperparameters, then adjusts the remaining one, and finally finds the hyperparameter with the best performance. The approach in this paper does not result in the best combination of parameters, but it can significantly reduce the hyperparameter optimization time. The results of the parameter sensitivity experiments for the multitask learning parameters λ_1, λ_2, λ_3 are shown in Figure 6. The experimental results show that in the case of temperature coefficient randomly fixed to 7, when λ_1 is 1 and λ_2 is 0.7, the F1 score of the model is the highest at 70.37 when λ_3 is chosen to be 0.5. Therefore, in this paper, we choose the λ_3

hyperparameter of pre-training multitasking learning to be 0.5. Similarly, in the case of a temperature coefficient randomly fixed at 7, when λ_1 is 1 and λ_3 is 0.5, the highest F1 score of the model is 70.41 when λ_2 is chosen as 0.5; when λ_2 is 0.5 and λ_3 is 0.5, the highest F1 score of the model is 0.8 when λ_1 is chosen as 0.8 is 71.09.

Figure 6. Multitask parameter sensitivity analysis.

For the pre-training temperature coefficient τ, the effect of the value of the pre-training temperature coefficient τ on the F1 score of the model is shown in Figure 7, with the parameters of λ_1, λ_2, λ_3 of the multitask learning fixed at 0.8, 0.5 and 0.5. The experimental results show that the F1 scores of the ABGFMIR model reach the peak at the values of temperature coefficient τ taken as 3, 7 and 10. In particular, the F1 score of the ABGFMIR model peaked at 71.09 in the case where the pre-training parameter temperature coefficient τ was taken as 7. Therefore, the final hyper-parameters were chosen as 0.8 for λ_1, 0.5 for λ_2, 0.5 for λ_3, and τ as 7.

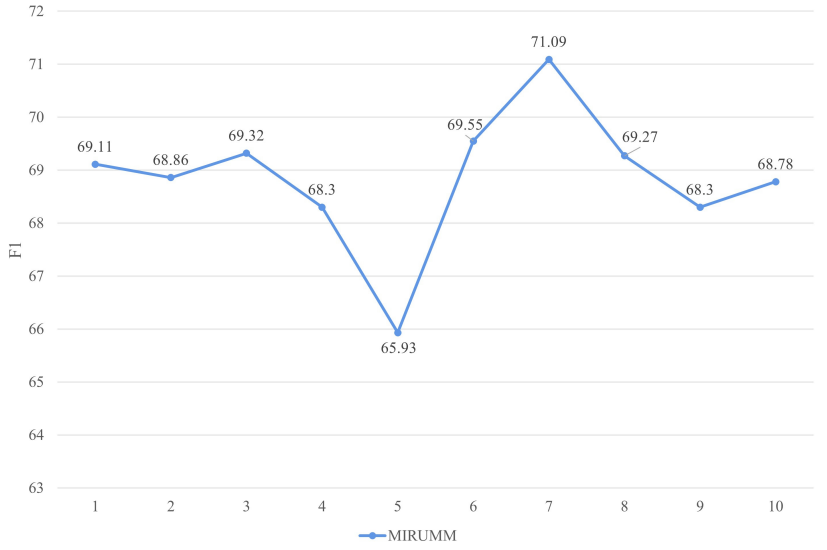

Figure 7. Sensitivity analysis of pre-training temperature coefficient parameters.

4.5. Performances Comparison

To verify the performance of ABGFMIR under uncertain missing modalities, we compare the accuracy and F1 scores of ABGFMIR with the baseline models through setting

different missing modalities rates between 0% and 50%. Experimental results are presented in Table 2. From Table 2 we can see that the F1 scores of each model show a decreasing trend with the increase in the missing rates. However, the accuracy of some models increases with the increase in missing rate; this is due to the fact that the missing samples are mainly concentrated in a few categories, and some models are more inclined to be predicted as most of the categories in order to increase the accuracy during the training process. As the F1 scores take into account both the precision rate and the recall rate, they may perform poorly on a few categories, thus leading to a decrease in F1 score.

Table 2. Experimental results comparing different missing uncertainty modal rates.

Dataset	Models	0%		10%		20%		30%		40%		50%	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MIntRec	MAG_Bert	72.65	68.64	68.41	66.12	65.23	63.75	63.86	62.03	61.59	60.59	60.23	56.51
	Mult	72.52	69.25	67.95	64.56	64.77	62.18	63.86	61.59	60.45	60.66	58.86	57.11
	Misa	72.29	69.32	68.41	65.65	63.64	61.67	62.95	59.44	56.82	56.33	58.86	55.75
	IF-MMIN	67.86	66.07	64.96	63.16	62.05	60.26	61.57	59.24	59.03	55.17	58.86	53.52
	EMRFM	72.58	70.46	68.41	66.81	62.05	65.95	63.18	63.14	60.00	59.00	57.95	58.18
	Ours	**73.86**	**71.09**	**70.00**	**67.68**	**66.14**	**66.20**	**65.91**	**64.48**	**62.73**	**62.27**	**61.82**	**60.12**
EMOTyDA	MAG_Bert	56.45	46.39	55.24	43.52	53.83	42.38	51.21	41.68	48.79	39.48	52.82	38.61
	Mult	59.48	46.36	55.44	44.03	54.03	41.64	54.64	41.10	51.81	40.62	50.00	38.57
	Misa	59.27	45.32	57.66	44.10	**56.05**	42.57	**55.65**	41.69	46.77	40.82	**57.86**	38.14
	IF-MMIN	53.83	35.58	54.44	34.47	52.62	33.83	50.00	29.26	44.76	27.76	48.19	25.82
	EMRFM	59.07	46.68	58.06	44.08	55.24	41.73	53.83	40.10	52.42	39.50	51.61	37.94
	Ours	**59.88**	**47.25**	**58.47**	**44.60**	55.65	**43.25**	55.04	**42.70**	**53.23**	**40.94**	53.63	**40.70**

From Table 2 we can see that on the dataset MIntRec, our proposed model ABGFMIR outperforms the other baseline models in terms of accuracy and F1 scores when the missing uncertainty modal rates are set to 0%, 10%, 20%, 30%, 40% and 50%. Compared with other baseline models, ABGFMIR improved accuracy by an average of 2.68 and improved F1 values by an average of 3.24.

On the dataset EMOTyDA, the ABGFMIR model outperforms the other baseline models in terms of accuracy and F1 scores when the uncertain modal missing rate is set to 0%, 10%, and 40%. When the modalities missing rate is set to 20%, the accuracy of the ABGFMIR model is 0.49 lower than that of the MAG_Bert model. When the modalities missing rate is set to 30%, the accuracy of the ABGFMIR model is only 0.61 lower than that of the MAG_Bert model. When the modalities missing rate is set to 50%, the accuracy of the ABGFMIR model is 4.23 lower than that of the MAG_Bert model. Although ABGFMIR does not perform as well as MAG_Bert in terms of accuracy under the above missing rates, the F1 score metrics of the ABGFMIR model outperform those of MAG_Bert for all the missing rates. In actuality, the F1 scores take into account the classification model's performance when dealing with the imbalances and concerns of misclassification more comprehensively than the accuracy metric, which is especially useful in multicategorization problems. Moreover, compared with other baseline models, ABGFMIR improved accuracy by an average of 2.28 and improved F1 values by an average of 3.44. In all, the above experimental results prove that ABGFMIR has better performance than that of other baseline models.

4.6. Ablation Experiment

To better verify the effectiveness of different modules in ABGFMIR, we conducted modal ablation experiments and module ablation experiments on dataset MIntRec. When conducting modal ablation experiments, single modal ablation experiments and dual modal ablation experiments are conducted separately. When conducting single mode ablation experiments, similarity learning is not performed after cross-modal fusion due to the use of only text as one modality. In addition, there is no uncertain mode missing in single modality, and only experiments with a missing rate of 0% are conducted in single modality.

The bimodal ablation experiment involves pairwise combinations of three modalities (text, visual and audio). For the module ablation experiment, variants of the ABGFMIR model are generated by removing different modules. Firstly, we remove the multimodal feature fusion module from ABGFMIR, and replace it with concatenation after extracting modality specific features. Then, we remove the pre-trained module (AGNN); finally, we remove the modality missing prompt module. Similar to the modal ablation experiment, we conducted ablation experiments with different modalities missing rates ranging from 0% to 50%.

Experimental results are presented in Table 3. Results of modal ablation experiments are presented as (1)–(7); from these results we can see that in multimodal intent recognition, the text modality has the strongest importance, and the visual and audio modalities only play a complementary role for the text modality. According to the results of modal ablation experiments, it was found that adding visual or audio modalities to text resulted in improved performance. Among them, the F1 score and accuracy of text plus audio are 3.18 and 2.03 higher than those of text plus visual when the missing rate is 0. This is because the audio modality may be more resistant to some environmental noise compared to the visual modality. However, the performance of audio plus visual has significantly decreased, which indicates that the text modality is the key modality for multimodal intent recognition. From these results, it can be seen that the text modality has significant advantages in multimodal intent recognition. In addition, the experimental results once again demonstrate the advantage of high correlation and complementarity between different modalities in multimodal intent recognition.

Table 3. Experimental results comparing different missing uncertainty modal rates.

Models	0%		10%		20%		30%		40%		50%	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
(1) T	70.00	66.14										
(2) V	16.82	3.29										
(3) A	16.14	8.02										
(4) T+A	70.91	68.21	70.68	66.80	67.27	65.44	63.37	63.96	60.46	60.14	55.68	57.34
(5) T+V	67.73	66.18	67.05	64.38	66.14	63.82	60.00	60.15	55.24	55.21	51.59	53.25
(6) A+V	17.50	12.39	18.18	9.26	18.64	9.27	15.45	8.01	17.50	7.87	15.68	5.41
(7) T+A+V	73.86	71.09	70.00	67.68	66.14	66.20	65.91	64.48	62.73	62.27	61.82	60.12
(8) -Fusion	73.18	69.10	69.09	66.82	65.00	64.73	62.73	61.08	61.59	59.55	60.68	58.28
(9) -Pre	72.27	70.23	69.55	65.93	63.86	64.59	63.64	62.43	62.05	60.12	60.23	59.88
(10) -Prompt	71.36	69.32	67.95	65.60	64.77	63.88	63.41	62.31	62.05	60.96	59.77	59.40

Specifically, under the condition of uncertain modal loss rates ranging from 0% to 50%, removing the multimodal feature fusion module resulted in a decrease in F1 score of 1.99, 0.86, 1.47, 3.4, 2.72, and 1.84, respectively. Removing the pre-training module resulted in a decrease of 0.86, 1.75, 1.61, 2.05, 2.15, and 0.24 in F1 scores, respectively. Removing the modal prompt module resulted in a decrease of 1.77, 2.26, 2.32, 2.17, 1.31, and 0.72 in F1 scores, respectively. When the missing rates are 0%, 30%, 40%, and 50%, removing the multimodal feature fusion module results in the greatest decrease in F1 score, indicating that the multimodal feature fusion module plays a dominant role in ABGFMIR.

4.7. Comparison of Different Grain Size Classifications

To analyze the performance of the ABGFMIR model for categorization with different granularities, we conduct experiments on the MIntRec dataset with different granularities. Experimental results are shown in Table 4. The coarse-grained categorization of the MIntRec dataset consists of two categories, “Expressing emotions or attitudes” and

"Achievement of goals", and the fine-grained contains 20 categories. Experimental results in Table 4 show that the F1 scores and accuracies of both coarse and fine-grained categorization decrease with the increase in uncertain modal missing rate. The changes of F1 scores and accuracy for different granularity classifications reveal that ABGFMIR achieves a substantial improvement when performing coarse-grained classification with fewer categories. Moreover, as the modalities missing rate increases, the F1 score and accuracy of ABGFMIR in coarse-grained classification decrease less compared to those of ABGFMIR in fine-grained classification.

Table 4. Experimental results comparing different missing uncertainty modal rates.

	Missing Rate	**Fine-Grained**		**Coarse-Grained**	
		ACC	**F1**	**ACC**	**F1**
(1)	0%	73.86	71.09	90.45	90.37
(2)	10%	70.00	67.68	**89.55**	**89.37**
(3)	20%	66.14	66.20	**88.64**	**88.46**
(4)	30%	65.91	64.48	**87.50**	**87.32**
(5)	40%	62.73	62.27	**87.27**	**87.01**
(6)	50%	61.82	60.12	**86.82**	**86.63**

Specifically, as the modalities missing rate of the MIntRec dataset increases from 0% to 50%, the F1 scores and accuracies of ABGFMIR in coarse-grained classification decrease by 3.74 and 3.63, with an average decrease of 0.75 and 0.73 each time, while those of ABGFMIR in fine-grained classification decrease by 10.94 and 12.04, with an average decrease of 2.19 and 2.41 each time. The above results are due to the F1 score and accuracy of coarse-grained classification being much more stable in the case of broader categories classification; the categories are broader and the model may still be able to classify relatively accurately with other available information for some samples that show uncertain modal deficiencies. Compared to coarse-grained categorization, multiple, more specific and detailed subcategories in fine-grained categorization cause confusion and lead to difficulty in model convergence. Based on the above experimental results, we can conclude that ABGFMIR has better robustness in coarse-grained classification compared that of ABGFMIR in fine-grained classification.

5. Conclusions

To solve the problem of multimodal intent recognition under uncertain missing modalities, we propose an Attention Bidirectional Gated Fusion Based Multimodal Intent Recognition model (named ABGFMIR). Firstly, to investigate the interaction of multimodal feature fusion, ABGFMIR adopts an attention bidirectional gated multimodal feature fusion module and a constraint strategy based on the CMD distance to bring the audio modality and visual modality closer to the textual modality. For the uncertain missing modalities, ABGFMIR directs the model's attention to these missing modalities by adding appropriate prompts in the Transformer. Moreover, we propose a pre-trained model which is trained with the complete modality to supervise the learning process of ABGFMIR at the classification layer, thus improving the generalization performance of ABGFMIR. Extensive experiments were conducted on two public benchmark datasets (MIntRec and EMOTyDA), and experimental results prove that our proposed model ABGFMIR has better performance in multimodal intent recognition under uncertain missing modalities than other baseline models. In our future work, we will conduct research on multimodal intent recognition driven by data and knowledge fusion.

Author Contributions: L.S.: Original draft, Visualization, Validation, Methodology; Z.L.: Methodology, Investigation, Formal analysis, Supervision; Y.W.: Review, Investigation, Validation, Software; X.S.: Methodology, Validation; J.Y.: Supervision, Editing; Q.Z.S.: Supervision, Validation, Formal analysis; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by: National Cultural and Tourism Technology Innovation Research and Development Project, National Natural Science Foundation of China (Grant nos. 62273290),Key Research and Development Project of Shandong Province (no. 2025CXPT077), Australian Research Council (ARC) Future Fellowship FT140101247 and Discovery Project DP200102298, and Yantai Institute of Science and Technology's Research Project YK-2025-1293.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data used in the experiments of our work are public benchmark datasets. For the first dataset MIntRec, the download link is https://github.com/thuiar/MIntRec (accessed on 1 November 2023); For the second dataset EMOTyDA, the download link is https://opendatalab.com/OpenDataLab/EMOTyDA (accessed on 1 November 2023).

Conflicts of Interest: Author Ling Shang is employed by the company Henan Culture and Tourism Investment Group Co. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Louvan, S.; Magnini, B. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In Proceedings of the 28th International Conference on Computational linguistics, Barcelona, Spain, 8–13 December 2020; pp. 480–496.
2. Singhal, B.; Gupta, A.; Shivasankaran, V.; Krishna, A. IntenDD: A unified contrastive learning approach for intent detection and discovery. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 14204–14216.
3. Sung, M.; Gung, J.; Mansimov, E.; Pappas, N.; Shu, R.; Romeo, S.; Zhang, Y.; Castelli, V. Pre-training intent-aware encoders for zero-and few-shot intent classification. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 10433–10442.
4. Zhang, F.; Chen, W.; Ding, F.; Gao, M.; Wang, T.; Yao, J.; Zheng, J. From Discrimination to Generation: Low-Resource Intent Detection with Language Model Instruction Tuning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 10174–10187.
5. Kim, M.; Choi, J.; Jang, C.; Lee, J. Bridging the Missing-Modality Gap: Improving Text-Only Calibration of Vision Language Models. *arXiv* **2026**, arXiv:2605.12517.
6. Zhu, Z.; Zhang, F.; Zhang, Y.; Sun, J.; Huang, Z.; Long, Q.; Xing, B.; Wu, X. A Survey on Multi-modal Intent Recognition: Recent Advances and New Frontiers. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 4–9 November 2025; pp. 15223–15236.
7. Li, H.; Yang, Q.; Xia, Y.; Lu, L.; Wei, Q. SaliText: A multimodal intent recognition method with saliency and text-guided fusion. *Signal Process.* **2026**, *244*, 110537. [CrossRef]
8. Gui, Q.; Liu, X.; Wang, J.; Ouyang, X.; Huang, W.; Zong, L. Evidence-driven ternary contrastive learning with hierarchical mamba fusion for robust multimodal intent recognition. *Neurocomputing* **2026**, *674*, 132866. [CrossRef]
9. Li, Y.; Wang, T.; Tang, M.; Hu, J. Text-guided frequency-decoupled modeling for multimodal intent recognition. *Neurocomputing* **2026**, *687*, 133795. [CrossRef]
10. Wang, M.; Xie, L.; Li, C.; Wang, X.; Sun, M.; Liu, Z. MGC: A modal mapping coupling and gate-driven contrastive learning approach for multimodal intent recognition. *Expert Syst. Appl.* **2025**, *281*, 127631. [CrossRef]
11. Sun, K.; Xie, Z.; Ye, M.; Zhang, H. Contextual Augmented Global Contrast for Multimodal Intent Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 26953–26963.
12. Huang, X.; Ma, T.; Jia, L.; Zhang, Y.; Rong, H.; Alnabhan, N. An effective multimodal representation and fusion method for multimodal intent recognition. *Neurocomputing* **2023**, *548*, 126373. [CrossRef]
13. Hu, B.; Zhang, K.; Zhang, Y.; Ye, Y. Adaptive multimodal fusion: Dynamic attention allocation for intent recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 17267–17275.

14. Wang, X.; Zhou, Y.; Huang, B.; Chen, H.; Zhu, W. Multi-modal Generative AI: Multi-modal LLMs, Diffusions and the Unification. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *36*, 5621–5641.
15. Zhang, C.; Cui, Y.; Han, Z.; Zhou, J.T.; Fu, H.; Hu, Q. Deep partial multi-view learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2402–2415.
16. Qin, L.; Xie, T.; Che, W.; Liu, T. A survey on spoken language understanding: Recent advances and new frontiers. *arXiv* **2021**, arXiv:2103.03095.
17. Qin, L.; Liu, T.; Che, W.; Kang, B.; Zhao, S.; Liu, T. A co-interactive transformer for joint slot filling and intent detection. In *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2021; pp. 8193–8197.
18. Wu, D.; Ding, L.; Lu, F.; Xie, J. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. *arXiv* **2020**, arXiv:2010.02693.
19. Qin, L.; Xu, X.; Che, W.; Liu, T. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. *arXiv* **2020**, arXiv:2004.10087.
20. Hou, Y.; Lai, Y.; Wu, Y.; Che, W.; Liu, T. Few-shot learning for multi-label intent detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 19–21 May 2021; pp. 13036–13044. [CrossRef]
21. Wu, T.W.; Su, R.; Juang, B. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 4884–4896.
22. Jia, M.; Wu, Z.; Reiter, A.; Cardie, C.; Belongie, S.; Lim, S.N. Intentonomy: A dataset and study towards human intent understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12986–12996.
23. Joo, J.; Li, W.; Steen, F.F.; Zhu, S.C. Visual persuasion: Inferring communicative intents of images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 216–223.
24. Fang, Z.; López, A.M. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4773–4783. [CrossRef]
25. Yang, B.; Zhan, W.; Wang, P.; Chan, C.; Cai, Y.; Wang, N. Crossing or not? Context-based recognition of pedestrian crossing intention in the urban environment. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 5338–5349. [CrossRef]
26. Xu, B.; Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, M.S. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Trans. Multimed.* **2019**, *22*, 1423–1432.
27. Zhang, H.; Xu, H.; Wang, X.; Zhou, Q.; Zhao, S.; Teng, J. Mintrec: A new dataset for multimodal intent recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 1688–1697.
28. Maharana, A.; Tran, Q.H.; Dernoncourt, F.; Yoon, S.; Bui, T.; Chang, W.; Bansal, M. Multimodal intent discovery from livestream videos. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 476–489.
29. Kruk, J.; Lubin, J.; Sikka, K.; Lin, X.; Jurafsky, D.; Divakaran, A. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv* **2019**, arXiv:1904.09073.
30. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2611–2624.
31. Liu, H.; Wang, W.; Li, H. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. *arXiv* **2022**, arXiv:2210.03501.
32. Zhang, L.; Shen, J.; Zhang, J.; Xu, J.; Li, Z.; Yao, Y.; Yu, L. Multimodal marketing intent analysis for effective targeted advertising. *IEEE Trans. Multimed.* **2021**, *24*, 1830–1843. [CrossRef]
33. Zeng, J.; Liu, T.; Zhou, J. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 1545–1554.
34. He, Y.; Zheng, S.; Tay, Y.; Gupta, J.; Du, Y.; Aribandi, V.; Zhao, Z.; Li, Y.; Chen, Z.; Metzler, D.; et al. Hyperprompt: Prompt-based task-conditioning of transformers. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 8678–8690.
35. Liu, Z.; Zhou, B.; Chu, D.; Sun, Y.; Meng, L. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Inf. Fusion* **2024**, *101*, 101973. [CrossRef]
36. Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; Poria, S. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv* **2019**, arXiv:1906.01815.
37. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.

38. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large Pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
39. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Online, 12–16 October 2020; pp. 1122–1131.
40. Zuo, H.; Liu, R.; Zhao, J.; Gao, G.; Li, H. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.