BiMIN: Bilateral Modality Imagination Network for Sentiment Analysis with Uncertain Modalities

Xuanchao Lin[a], Junjie Peng[a,b,*]

[a]School of Computer Engineering and Science, Shanghai University, Shanghai, China
[b]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China

ARTICLE INFO

Keywords:
Multimodal Sentiment Analysis
Modality Reconstruction
Unified Model

ABSTRACT

In previous studies, Multimodal Sentiment Analysis(MSA) using textual, acoustic and visual modality has been proven to improve the performance of Sentiment Analysis. However, in reality, it is not always possible to obtain the three modalities. In different reality scenes, only one or two modalities can be obtained, while the remaining are absent. Model of Sentiment Analysis for the certain modalities will be ineffective in different situations. In view of the possibility of the absence of the three modalities, the focus of the previous work is on the reconstruction of the absent modalities by using the given available modalities. Therefore, we propose the Bilateral Modality Imagination Network (BiMIN) as a unified model to address various scenarios of absent modalities. This network incorporates two key mechanisms: 1) Pre-trained Encoders to improve the usability of acoustic and visual modalities. 2) Bilateral Imagination Module to improve the performance of reconstruction of absent modalities. Comprehensive experiments on three benchmark datasets (**CMU-MOSI**, **CMU-MOSEI**, **CH-SIMS**) show that the BiMIN model significantly improves sentiment analysis performance under various combinations of available modality test conditions.

1. Introduction

Human beings receive information from the surroundings through a variety of senses to construct their understanding of the world. The world is filled with different forms of information like text, sound and image which are the main information transmission media. There, sentiment analysis is mainly used to identify a person's actual sentiment tendency through textual modality, acoustic modality and visual modality, or two or all of them.

In real life, the types of modality available to people in different situations is not fixed. For example, when talking to someone on the phone, only the voice and the corresponding text information are available; When communicating with a foreigner who does not speak the same language, only his facial expression, movement, voice, etc. are available, and there is no comprehensible text of what he's saying; When communicating with a patient who is temporarily aphasic but able to express himself by writing or typewriting, only the text and the visual information are available. Situations similar to the above are common in reality, where only two or even only one modality is available in the communication process, as shown in the Table 1. General models for sentiment analysis are based on ideal fixed scenarios, such as Tensor Fusion Network (TFN) [1] and Multimodal Transformer (MulT) [2] for tri-modality, Bi-directional Modality Fusion Network (BMFN) [3] for bimodality of audio-video, Lifelong Text-Audio Sentiment Analysis (LTASA) [4] for bimodality of text-aduio, Deep Coupled Video and Danmu Neural networks (DCVDN) [5] for bimodality of video-text, Knowledge-Enriched Attention-based Hybrid Transformer (KEAHT) [6] for unimodality of textual modality, the parallel Network for multi-scale SER based on a connection Attention Mechanism (AMSNet) [7] for unimodality of acoustic modality, and Clip-aware Emotion-rich Feature Learning Network (CEFLNet) [8] for unimodality of visual modality. In response to varying real-world situations mentioned above, the network may need to be remodeled accordingly. For example, if a multimodal sentiment analysis model for tri-modality is uesd in bimodality available situations, it will fail due to the absence of some modalities. Faced with this task of sentiment analysis with uncertain modalities, the key is usually twofold: on the one hand, to mine as much sentiment information as possible for the available modalities, and on the other hand, to complete the generation of absent modalities for better sentiment analysis.

Table 1
The six possible absent-modality combinations."$\surd$" denotes the modality is available.

	{Available}	Text	Vision	Audio	Absent Combinations
1	$\{a\}$			$\surd$	$(x^t_{absent}, x^v_{absent}, x^a)$
2	$\{v\}$		$\surd$		$(x^t_{absent}, x^v, x^a_{absent})$
3	$\{t\}$	$\surd$			$(x^t, x^v_{absent}, x^a_{absent})$
4	$\{a, v\}$		$\surd$	$\surd$	(x^t_{absent}, x^v, x^a)
5	$\{v, t\}$	$\surd$	$\surd$		(x^t, x^v, x^a_{absent})
6	$\{a, t\}$	$\surd$		$\surd$	(x^t, x^v_{absent}, x^a)

mine sentiment information for the available modalities. Several studies [1, 9–11] and corresponding ablation experiments, have shown that there are significant differences in the contribution of different modalities to sentiment analysis, with textual modality dominating. Our experiments in Section 4.4 also show that the performance decreases significantly when textual modality is absent, which further confirms the dominant role of textual modality in sentiment

*Corresponding author
linxuanchao@shu.edu.cn (X. Lin); jjie.peng@shu.edu.cn (J. Peng)

analysis. Because the non-textual modalities is inferior to the textual modality, many models underperformed in scenes, that lack textual modality. Therefore, how to effectively mine sentiment information in non-textual modalities has become an important problem to be solved.

complete the generation of absent modalities. The mainstream solution is to use Auto-Encoder (AE) to complete the generation of absent modalities. Cascaded Residual Auto-Encoder (CRA) [12] cascades multiple Auto-Encoders together in the way of residual networks (As shown in Figure 1), which is able to construct joint multimodal representations for multimodal sentiment analysis while generating absent modalities. Models such as Missing Modality Imagination Network (MMIN) [13] and Invariant Features for a Missing Modality Imagination Network(If-MMIN) [14] are based on this method were further researched and developed. In the specific prediction process, a joint multimodal representation is constructed by integrating the Hidden States generated by each Auto-Encoder, and this representation is then used for sentiment prediction. The encoder part of the Auto-Encoder follows a single-branch design, where the data dimensions are progressively reduced during forward propagation, inevitably leading to information compression. The hidden states of the Auto-Encoder can only retain core information, which causes the loss of detailed information. Although a residual connection supplements the original information after the Auto-Encoder's forward output, the final input for sentiment prediction, composed of the hidden states, still lacks detailed information. This result in performance bottlenecks in sentiment prediction. Therefore, exploring effective strategies to address the performance bottlenecks caused by the single-branch design of Auto-Encoder has become particularly crucial.

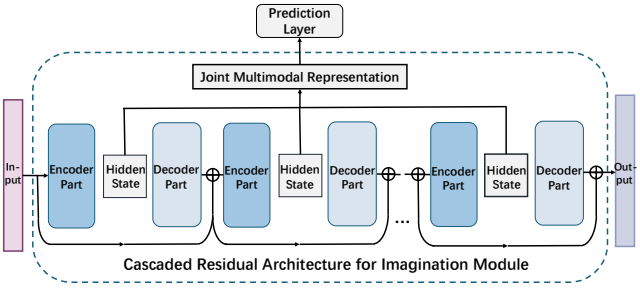


Fig. 1. Illustration of the Cascaded Residual Architecture for Imagination Module. The Encoder Part and Decoder Part as a unified entity, which can either be an **Original Auto-Encoder** or a **Bilateral Auto-Encoder** designed by us.

To solve the above problems, we propose **Bilateral Modality Imagination Network (BiMIN)**, which serves as a unified framework capable of handling sentiment analysis under modality uncertainty. In addition to managing the six possible absent-modality combinations, BiMIN is also able to perform sentiment analysis effectively when all three modalities are available. BiMIN is able to dig deeper into the sentiment information of each available modality and enhance the robustness of the joint multimodal representations to significantly improve the performance of sentiment analysis. architecture of BiMIN is composed of two core components: the first is the **Modality Encoder Network** for feature extraction and coding of different modality features; The second is the **Bilateral Imagination Module** for generation of absent modalities and further enhancement of the interaction and integration between modalities.

The **Modality Encoder Network** was equipped with pre-trained models to extract features from different modalities. As mentioned before, in sentiment analysis research, the textual modality plays the most critical role, as it contains more semantic information compared to the audio or visual modalities. the high performance of the textual modality can be attributed to two main factors. First, the high abstractness of the text gives it a unique advantage in sentiment analysis, because certain words (e.g. *like*, *hate*, etc.) are strongly correlated with sentiment states. Second, the extraction of textual modality feature primarily relies on large-scale pre-trained models such as BERT [15], which can deeply explore advanced semantic features within the text, significantly enhancing the performance of sentiment analysis tasks. However, for non-text modalities, traditional studies typically employs extraction tools to obtain various sub-features related to sentiment analysis from raw audio and video data. Although these sub-features are structurally distinct and interrelated, the direct concatenation of them as non-textual modality features, which makes it difficult for feature encoders to effectively recognize and integrate these diverse sub-features, resulting in performance bottlenecks [16].We drew inspiration from the feature extraction strategy used for the textual modality, applying pre-trained models to extract features from non-textual modalities. This approach aims to eliminate performance bottlenecks caused by different sub-feature types and to extract features rich in high-level semantic information. Even in the absence of textual information, our model is capable of maintaining robust performance.

the **Bilateral Imagination Module** consists of multiple **Bilateral Auto-Encoders** integrated through a cascading residual connection approach, as shown in Figure 1. The Bilateral Auto-Encoder we designed aim to mitigate the loss of detailed information that typically occur during the encoding process in traditional Auto-Encoder. As mentioned earlier, the performance bottleneck in sentiment analysis caused by the loss of detailed information stems from the single-branch design of traditional Auto-Encoder. In contrast, the Bilateral Auto-Encoder we designed primarily relies on two independent yet cooperative branches: the Core Branch is responsible for encoding and extracting advanced semantic features, while the Peripheral Branch focuses on preserving critical detailed information. This dual-branch architecture effectively mitigates the issue of detail loss inherent in single-branch designs, thereby improving sentiment analysis performance.

Experimental results on three public datasets for MSA show that the method generally outperforms the benchmark models under all available modality combination conditions.

The main contributions of this work are:

- We design a Bilateral Modality Imagination Network aimed at addressing sentiment analysis under conditions of modality uncertainty.
- Our Modality Encoder Network leverages pre-trained models to extract features from non-text modalities, thereby enhancing the usability and expressiveness of non-textual data.
- The Bilateral Autoencoder we designed is capable of simultaneously capturing and preserving both core and detailed information of the modalities, forming a Bilateral Imagination Module that enhances the robustness of joint multimodal representations.
- Extensive experimental results on three major multimodal sentiment analysis datasets demonstrate that our model significantly improves sentiment analysis performance under varying conditions of modality availability.

2. Related work

2.1. Certain Modalities Sentiment Work

Multimodal Sentiment Analysis focuses on the process of analysing and understanding human sentiments using multiple forms of data such as text, audio and video. The main challenge of this task is to effectively use information from different modalities to complement each other in order to predict the subject's sentiments. However, variations in scenarios lead to differences in available modalities, and these differences necessitate distinct model solutions and frameworks tailored to the specific combinations of modalities.

For the bimodality work. BMFN [3] focuses on **audio-video bimodality** fusion, proposing an mechanism that enhances the representation of acoustic and visual features through bidirectional fusion, while refining and adjusting these features using a feedforward-backward attention module. LTASA [4] is centered on **audio-text bimodality**, which designs a complementary-aware subspace specifically to explore the nonlinear complementary knowledge between text and audio modalities. For **video-text bimodality** analysis, DCVDN [5] integrates video content with real-time text, utilizing deep learning techniques to synchronously extract and fuse visual and textual features, and employs a Deep Canonically Correlated Auto-Encoder for multi-view learning, enabling precise analysis of audience sentiment in videos.

For the tri-modality work. TFN [1] models the relationship between different modalities by computing the Cartesian product. MulT [2] align the different modalities by using cross-modality attention mechanism, addressing the challenges of modality misalignment and heterogeneous information fusion. Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) [17] designs a Unimodal Label Generation Module based on a self-supervised learning strategy to acquire independent unimodal supervision and jointly trains on multimodal and unimodal tasks. Modality-Invariant and - Specific Representations for Multimodal Sentiment Analysis (MISA) [9] projects each modality into two different subspaces and learns modality invariance as well as modality specificity respectively. Recurrent Attended Variation Embedding Network (RAVEN) [18] models the fine-grained structure of non-textual sequences and dynamically shifts word representations based on non-textual cues.

The aforementioned works have all demonstrated strong performance in multimodal sentiment analysis tasks. However, in real-world scenarios, available modalities is often uncertain, which introduces new challenges to sentiment analysis tasks.

2.2. Uncertain Modalities Sentiment work

Due to the uncertainty of scenarios in reality, the absence of certain modalities can cause sentiment analysis models designed for fixed-modality scenarios to fail. To address this issue, previous research has primarily focused on reconstructing the missing modalities using the available ones and combining them to form joint multimodal representations for effective multimodal sentiment analysis.

Based on cross-modality attention. Multimodal Cyclic Translation Network (MCTN) [19] uses a sequence-to-sequence model to perform cyclic translation between different modalities, and is robust to noise and missing data, and is able to learn robust joint representations. Transformer-based Feature Reconstruction Network (TFR-Net) [20] is a feature reconstruction network based on Transformer [21], designed to handle randomly missing data in unaligned modality sequences. It extracts robust representations for each modality sequence through internal and cross-modality attention mechanisms. Coupled-Translation Fusion Network (CTFN) [22] utilises a hierarchical structure of multiple bi-directional translations to achieve dual multimodal fusion embedding compared to MCTN. Although cross-modality attention-based methods effectively capture inter-modality relationships, they are computationally intensive, lack real-time performance, and underperform when a modality is significantly missing or absent.

Based on Auto-Encoder. Auto-Encoder-based generative methods learn to generate new data with similar distributions to the observed data. Due to their ability to learn latent representations, Auto-Encoder have been widely applied in multimodal sentiment analysis. CRA [12] is designed for data with missing modalities, combining a series of residual Auto-Encoders into a cascading architecture to learn the relationships between different modalities. MMIN [13] leverages CRA and Cycle Consistency Learning [23] to build robust joint multimodal representations, enabling the prediction of missing modality representations. Compared to CRA,

MMIN's unified framework is more adaptable, allowing stable operation under various missing conditions of modality. IFMMIN [14]extends MMIN by utilizing central moment discrepancy distance constraints to learn modality-invariant features, addressing the challenge of missing modalities in practical applications. Although these Auto-Encoder-based generative methods achieve good performance in sentiment analysis tasks with uncertain modalities, they still face performance bottlenecks due to the inherent design limitations of Auto-Encoder.

Based on Variational Auto-Encoder. Research based on Variational Auto-Encoder (VAE) [24] incorporates variational inference into Auto-Encoder, enhancing capabilities in data generation. VAE maps input data into a low-dimensional latent space via an encoder and reconstructs the data from this space using a decoder, simultaneously optimizing the reconstruction quality and the distribution of the latent representations to achieve efficient data generation and feature extraction. Conditional MultiModal Autoencoder (CMMA) [25] learns the conditional distribution between modalities and optimizes the conditional log-likelihood using variational inference methods, enabling the generation of one modality based on the condition of another, making it particularly suitable for multimodal tasks requiring conditional generation. Multi-view Variational Autoencoder (MVAE) [26] models the statistical relationships in multimodal sentiment data through multiple modality-specific generative networks and a shared latent space. By imposing a Gaussian mixture assumption on the posterior approximation of the shared latent variables, MVAE learns joint deep representations across modalities and assesses the importance of each modality. While models based on VAE offer advantages in generating new samples and inferring latent variables compared to traditional Auto-Encoder, they may face challenges such as high computational complexity, training difficulty, and the limitations imposed by the Gaussian assumption on the latent space distribution.

Optimization Strategies. Recent works have introduced optimization strategies based on the aforementioned approaches. Ensemble-based Missing Modality Reconstruction (EMMR) [27] integrates Auto-Encoder-based and cross-modality-attention-based methods to enhance decision-making capabilities through ensemble techniques. Multimodal Prompting with Missing Modalities (MPMM) [28] and Multimodal Prompt Learning with Missing Modalities (MPLMM) [29] optimize performance through lightweight methods by incorporating prompt information. Tag-Assisted Transformer Encoder (TATE) [30] designs a tag encoding module to mark missing modalities, guiding the network to focus on these absent data, and employs a new spatial projection method to align shared feature vectors.

Compared to previous work, we have designed a Bilateral Auto-Encoder based on the Auto-Encoder framework and developed the BiMIN. This model strikes a balance between accuracy and efficiency, effectively addressing the challenges of sentiment analysis in scenarios with uncertain modalities. It significantly improves performance, particularly in cases where the text modality is absent, with notable enhancements in those conditions.

3. Method

In this section, we first define the problem and then describe our BiMIN.

3.1. Task Setup

For a given set of video clips C, we represent the raw information from the three modalities of each clip $c \in C$ as $x = (x^t, x^v, x^a)$, where x^t represents the textual content of the clip, x^v represents the original video without sound, and x^a represents the audio in the clip. $|C|$ indicates the total number of video clips. Our target set is $Y = [y_i]_{I=1}^{|C|}$, where y_i denotes the target sentiment polarity of a video clip c_i. The method we propose aims to identify the sentiment polarity y_i for each video clip c_i, even when one or two modalities are absent. As illustrated in Figure 2, when the video modality is absent, only the acoustic and textual modalities are available for analysis.

3.2. Overall Framework

The proposed BiMIN framework as shown in figure 2, consists of three main components:

Modality Encoder Network: Different pre-trained models are first employed to extract the features of each modality from the raw video clip data. These features are then further processed by the modality encoder to obtain the initial modality embeddings h.

Bilateral Imagination Module: Given the initial embeddings of the available modalities h', the Bilateral Imagination Module generates imagined multimodal embeddings $\hat{h}$. During this process, Hidden States S_i from each Bilateral Auto-Encoder B_i are collected to form a joint multimodal representation S, where $i \in [1, M]$.

Sentiment Prediction Layer: Finally, the joint multimodal representation S is utilized to predict the sentiment polarity $\hat{y}$ of the video clips.

The following subsections will provide a detailed explanation of each part.

3.3. Modality Encoder Network

For a given raw video, we can extract three different unimodal sequences, represented as x^t, x^v and x^a, corresponding to the textual, visual, and acoustic modalities, respectively.

Textual Feature Extractor. We utilize a pretrained BERT [15] language model as the textual feature extractor. BERT is a self-supervised Transformer model that is pre-trained on massive corpus to learn inner representations of text with rich sentiment information. Although the BERT model architecture is based on a Transformer encoder-decoder during training, only the encoder part is used for feature extraction. Given the original sentence $x^t = \{w_1, w_2, ...w_n\}$ consisting of n words, we concatenate it with two special tokens [CLS] and [SEP], then input this

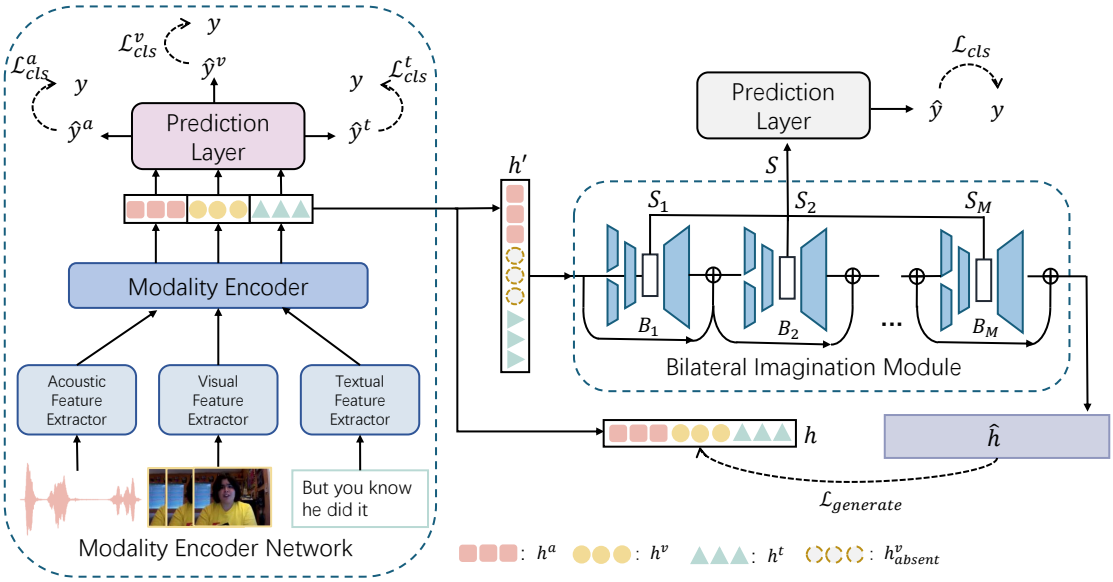


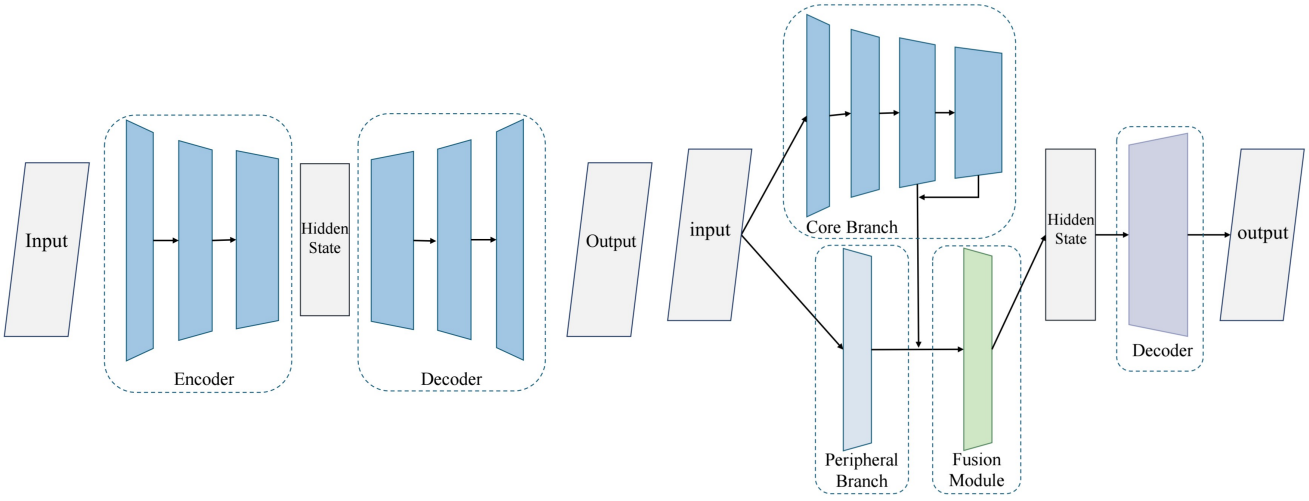
Fig. 2. Illustration of the Bilateral Modality Imagination (BiMIN) framework.

sequence into BERT to obtain the textual modality features F^t.

Visual feature Extractor. In the visual modality, previous works commonly employed Multi-task Cascaded Convolutional Networks (MTCNN) [31] or OpenFace [32][33] to capture facial expression features, including sub-features such as eye gaze direction, head pose, and facial action units. Although these sub-features have distinct structures, they are interrelated and collectively provide an objective description of facial expressions. However, these sub-features were often directly concatenated to form visual feature, and their intrinsic structure was not adequately learned by the models, which limited the performance of the final sentiment analysis tasks. To address this issue, we drew inspiration from text feature extraction methods and employed the pre-trained visual model X-CLIP [34] as visual feature extractor to capture the raw features from videos. X-CLIP is a language-image pre-training model for general video recognition, with its video encoder allowing information exchange between video frames and integrating frame-level representations to output video features. Specifically, we uniformly sampled 8 frames from the original video x^v and converted these frames into the required tensor format using the preprocessing functions provided by X-CLIP. The processed data was then fed into X-CLIP to extract the visual features F^v.

Acoustic Feature Extractor. For the acoustic modality, previous research typically used tools like OpenSMILE[25], COVAREP[26], and LibROSA[27] to extract acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), pitch, intensity, and peak frequency. These sub-features, though structurally distinct, are interrelated and collectively represent the emotional information in speech. However, similar to the visual modality, these acoustic sub-features are often directly concatenated as input to the model. This makes it challenging for the model to effectively recognize and learn from the heterogeneous data, resulting in performance bottlenecks. To address this issue, we utilized Whisper [35], a pre-trained large model, to extract audio features. Whisper is a pre-trained model designed for automatic speech recognition and speech translation, based on a Transformer encoder-decoder architecture and trained on a vast dataset of unlabeled speech, providing strong speech processing capabilities. Its encoder is capable of learning high-quality acoustic representations. Specifically, we transformed the raw audio data x^a using a Mel filter to convert it into a spectrogram, applying a logarithmic scale to enhance the details in the high-frequency range. The transformed data was then padded or trimmed as needed to fit the fixed input size required by Whisper. Finally, the processed data was fed into encoder of Whisper to obtain the acoustic modality features F^a.

Modality Encoder: We adopted a stagewise training approach to process the extracted text features F^t, visual features F^v, and acoustic features F^a. Specifically, we employed recurrent neural networks (RNNs) to capture the temporal relationships among these features and to unify their feature dimensions, resulting in textual embeddings h^t, visual embeddings h^v, and acoustic embeddings h^a. These embedding are concatenated to form the initial full-modality embedding $h = (h^t, h^v, h^a)$. For scenarios with varying numbers of available modalities, we represent a



(a) Original Auto-Encoder (b) Bilateral Auto-Encoder

Fig. 3. Architectures of two different Auto-Encoder

missing modality embedding h^m as a zero vector $h^m_{abesent}$ to indicate its absence, where $m \in \{t, v, a\}$. For example, if only the textual and acoustic modalities are available and the visual modality is absent, the corresponding initial embedding for the bilateral imagination module would be $h' = (h^t, h^v_{absent}, h^a)$. In the subsequent section 4.2, we may use a unified input format as (x^t, x_absent^v, x^a). Since our approach involves stepwise training, there is no substantive difference between the two formats for the Bilateral Imagination Network. This will not be explained further in later sections, but it is mentioned here to prevent any misunderstandings.

3.4. Bilateral Imagination Module

Auto-Encoder was initially developed as an unsupervised learning method, aimed at learning latent representations of data that can be used to reconstruct the original input. It encodes the input vector into a hidden representation through a nonlinear mapping, and then decodes it back through another mapping, as illustrated in Figure 3a. As the method is illustrated in Figure 1, The current CRA network connects Auto-Encoders in a cascaded residual Architecture, extracting the hidden state from each Auto-Encoder to form a joint multimodal representation for predicting sentiment polarity. Due to the single-branch of the original Auto-Encoder, its channel capacity is relatively low, inevitably leading to information compression. As a result, it can only retain some of the core information, causing the encoded hidden state to lose detailed information. Although a residual connection is added after the decoder's output to replenish the original information when entering the next Auto-Encoder, the joint multimodal representation formed by the hidden states fails to recover the lost detailed information. Consequently, this lack of detailed information in sentiment prediction leads to a bottleneck in the final prediction performance.

To address this issue, we propose the **Bilateral Auto-Encoder**, as illustrated in Figure 3b. The Bilateral Auto-Encoder consists of a peripheral branch and a core branch, designed to capture these two different types of information and thereby reduce the performance loss caused by the loss of detailed information.

The **Peripheral Branch** is responsible for retaining features that might be discarded during forward propagation but are still potentially useful. Similar to computer vision tasks, where edge, texture, and other fine details may be discarded during image processing, the Peripheral Branch requires substantial channel capacity to encode these detailed information of different modalities. Since it focuses primarily on low-level details, we employ wide channels and a shallow neural network in this branch to preserve these details.

The **Core Branch**, Running in parallel to the Peripheral Branch, connects multiple linear layers to capture deeper, richer, and more abstract semantics. A global average pooling layer is added at the end to provide a receptive field with global contextual information. The Core Branch has a lower channel capacity, which retains only the most essential semantics, while the Peripheral Branch complements this by preserving the detailed information that might be ignored.

The **Fusion Module** combines the information from both branches. The need for a fusion module arises primarily because the feature representations of the peripheral and core branches are complementary, with each branch not directly accessing the information from the other. Various methods can be used for information fusion, such as simple summation, self-attention mechanisms, and cross-modality attention mechanisms. After considering both accuracy and efficiency, we chose a simple linear mapping method for fusion. Specifically, we concatenate the information from both branches and then use the Fusion Module to map it into a Hidden State of a fixed dimension.

For the **Decoder**, we similarly considered both accuracy and efficiency, discarding the typical three-layer linear design used in original Auto-Encoder (as shown in Figure 3a) and opting instead for a single-layer linear decoder.

As shown in Figure 2, Bilateral Imagination Module is composed of M Bilateral Auto-Encoders. These Bilateral Auto-Encoders are connected in a cascaded residual architecture, as illustrated in Figure 1. The architecture of each Bilateral Auto-Encoder B_i is depicted in Figure 3b, where the core branch is denoted as cb_i, the peripheral branch as pb_i, the fusion module as fm_i, and the decoder as Dec_i, where $i \in \{1, 2, ..., M\}$. The computation for each Bilateral Auto-Encoder can be defined as:

$$\Delta_{Z_i} = \begin{cases} Dec_i\left(fm_i(cb_i(h'), pb_i(h'))\right), & \text{if } i = 1; \\ Dec_i\left(fm_i(cb_i(h' + \sum_{j=1}^{i-1}\Delta_{Z_j}), pb_i(h' + \sum_{j=1}^{i-1}\Delta_{Z_j}))\right), & \text{if } i > 1. \end{cases} \tag{1}$$

Here, h' represents the initial embedding composed of the available modalities, and Δ_{Z_i} denotes the output of the i^{th}bilateral Auto-Encoder. For the case where the visual modality is absent (as shown in Figure 2), the Modality Imagination Module generates a multimodal embedding based on the available acoustic and textual modalities. The multimodal embedding representation $\hat{h}$ generated by the Bilateral Imagination Module is:

$$\hat{h} = BiIM(h') = h' + \Delta_{Z_M} \tag{2}$$

where $BiIM(\cdot)$ representations the function of the Bilateral Imagination Module

We concatenate the hidden states S_i from each Bilateral Auto-Encoder B_i in the Bilateral Imagination Module to form a joint multimodal representation $S = (S_1, S_2, ..., S_M)$, which is used for subsequent sentiment analysis.

3.5. Prediction Layer

As illustrated in Figure 2, it is noticeable that there is an sentiment prediction layer within the Modality Encoder Network. Additionally, another sentiment prediction layer is connected outside the Bilateral Imagination Module.

The former corresponds to a sentiment prediction layer PL_m for each modality. The embedding of each modality h^m ,as mentioned in 3.3, is fed into its corresponding sentiment Prediction Layer PL_m to obtain the sentiment polarity prediction:

$$\hat{y}^m = PL_m(h^m) \tag{3}$$

where, $PL_m(\cdot)$ denotes the function of sentiment prediction layer for each modality m, and $m \in \{t, v, a\}$. This process is primarily used to train each Modality Encoder.

The latter is a sentiment prediction layer PL based on the joint multimodal representation S, used for the final overall sentiment polarity prediction:

$$\hat{y} = PL(S) \tag{4}$$

where $PL(\cdot)$ denotes the function of sentiment prediction layer for the joint multimodal representation S.

These sentiment prediction layers consist of multiple fully connected layers.

3.6. Loss Function

In the process of BiMIN training, we adopt a stagewise training method.

In the Modality Encoder Network, we train each modality-specific encoder within its respective modality. The training process of the unimodal initial embeddings is supervised by the classification loss $\mathcal{L}_{cls}^m$, which is defined as follows:

$$\mathcal{L}_{cls}^m = -\frac{1}{|C|}\sum_{i=1}^{|C|} H(y, \hat{y}^m) \tag{5}$$

where $H(y, \hat{y}^m)$ is the cross-entropy between distributions y and $\hat{y}^m$, and $m \in \{t, v, a\}$.

In the Bilateral Imagination Network, two primary loss functions guide the training. First, the classification loss $\mathcal{L}_{cls}$ is used to supervise the training process with sentiment polarity targets. Secondly, the generative loss $\mathcal{L}_{generate}$ supervises the training of the generated multimodal embeddings h':

$$\begin{aligned} \mathcal{L}_{cls}^m &= -\frac{1}{|C|}\sum_{i=1}^{|C|} H(y, \hat{y}) \\ \mathcal{L}_{generate} &= \|h; h'\|_2^2 \end{aligned} \tag{6}$$

Here, $H(y, \hat{y})$ is the cross-entropy function as same as mentioned before, and $\|\cdot\|_2^2$ is the squared Frobenius norm.The total loss $\mathcal{L}$ is the sum of these two functions:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{generate} \tag{7}$$

Where λ is a hyperparameter.

4. Experiments

4.1. Datasets and Evaluation Metrics

To replicate real-world conditions, we assessed our method using three well-established datasets for multimodal sentiment analysis: CMU-MOSI [36], CMU-MOSEI[37], and CH-SIMS [38].

The CMU-MOSI dataset is widely used for multimodal sentiment analysis. It includes 93 English YouTube videos, each clip annotated with sentiment scores on a scale from strongly negative to strongly positive (-3 to +3).

The CMU-MOSEI dataset expands on CMU-MOSI, offering more than 65 hours of annotated video content from over 1,000 speakers across 250 different topics. This dataset provides a wider coverage of topics compared to CMU-MOSI.

The CH-SIMS dataset is a Chinese multimodal sentiment analysis dataset, comprising 2,281 video clips. Each clip is annotated with sentiment scores ranging from strongly negative to strongly positive (-1 to 1).

In line with previous research, we use binary accuracy (ACC) and F1 score (F1) as evaluation metrics for both CMU-MOSI and CH-SIMS. For CMU-MOSEI, the evaluation metrics include 2-class accuracy (ACC), F1 score (F1), 5-class accuracy (ACC5), 7-class accuracy (ACC7), and Mean Absolute Error (MAE).

4.2. Experimental Setup

Original Feature Extraction. As a baseline comparison, we extracted frame-level raw features for each modality across all three datasets using the same method as [38].

Training and Test Set. We first define the original dataset, which includes all three modalities, as the full-modality dataset. Based on this full-modality dataset, we construct six different absent-modality combinations (as shown in Table 1) to simulate various possible scenarios of modality absence. For the training set, we combine the training subsets of these six different absent-modality combinations to create a comprehensive training set. For the test set, we consider each of the six absent-modality combinations as separate test subsets, corresponding to six conditions of modality absence. For instance, during the inference phase, if the visual modality is absent as shown in Figure 2, the unified format of the available modality test samples would consist of the original features (x^t, x^v_{absent}, x^a). These six test subsets represent the six modality absence scenarios and are used as the test sets.

Model Training Details. In all experiments, we employed the Adam optimizer [39] with a batch size of 64. The model was trained with an initial learning rate of 1×10^{-3} using early stopping with a patience of 8 epochs. To ensure reproducibility, we fixed the random seed so that each model was trained on the same data. All models were implemented using the PyTorch deep learning framework and were run on a single Nvidia GTX 4060Ti GPU.

4.3. Baseline

We compare our proposed model to the following:

- **MCTN**: Multimodal Cyclic Translation Network (MCTN, 2019) learns robust joint representations through modality translation to process missing information. [19]
- **TFR-Net**: Transformer-based Feature Reconstruction Network (TFR-Net, 2021) used Transform architecture to supplement missing information for emotion prediction. [20]
- **MMIN**: Missing Modality Imagination Network (MMIN, 2021) learns robust joint multimodal representations that can predict the representation of any missing mode given the available modes.[13]
- **If-MMIN**: Invariant Features for a Missing Modality Imagination Network (If-MMIN, 2022) extracts invariant features between different modalities in order to learn robust joint multimodal representations that can predict representations of any missing modalities given available modalities. [14]
- **MPMM**: Multimodal Prompting with Missing Modalities (MPMM, 2023) uses deletion-aware cues to guide models in solving missing modalities problems. [28]
- **MPLMM**: Multimodal Prompt Learning with Missing Modalities (MPLMM, 2024) uses a variety of cues to guide the generation of missing modalities features. [29]

4.4. Main Results

Comparison of the individual models of modality under different absent-modality combinations in Table 2, we present the quantitative results for the three datasets. Analysing the results given in the table, it evident that our proposed method consistently outperforms the baseline in all datasets for all six available-modality conditions.

When the textual modality is absent, which means the result of $\{a\}, \{v\}, and \{a, v\}$), our method delivers substantial improvements with accuracy increases of 7-12% compared to the best results from the baseline model. And when the textual modality is available ($\{t\}, \{a, t\}, \{v, t\}$), our model generally achieves superior performance.

For the **MOSI** dataset, while our model does not outperform TRF-Net and MPLMM under the specific $\{v, t\}$ condition, it still demonstrates superior overall performance. The average ACC and F1 score of our model exceed the baselines by 4%, highlighting its better generalization capabilities. Moreover, under the $\{v, t\}$ condition across all other datasets, our model consistently achieves the highest performance relative to the current baselines.

On the **SIMS** dataset, it was observed that under the $\{a\}$ and $\{v\}$ available-modality conditions, our model slightly slightly underperforms MPLMM and MPPM in terms of F1 scores, but our model surpasses them in terms of ACC. This suggests that although MPLMM and MPPM may excel in recall for certain classes, this comes at the cost of precision for the majority class. In contrast, our model not only maintains higher accuracy but also achieves similarly high F1 scores, underscoring its superior generalization ability. It is also noteworthy that SIMS is a Chinese dataset, and our model's strong performance on this dataset further demonstrates its effectiveness not only in English but also in Chinese contexts, highlighting its cross-linguistic generalization capability.

As for the **MOSEI** dataset, it is known for its large sample size. On this large dataset, our model shows exceptional performance across all six test cases, with significant improvements in both ACC and F1 scores compared to the baseline. Notably, for different modalities combinations, our model exhibits minimal fluctuations in performance, with deviations from the overall average performance kept at around 3%. This result underscores our model's robust handling of data incompleteness, maintaining stable performance even under challenging conditions. Additionally, our model not only excels in generalization but also adapts well to varying data environments, providing consistent and reliable performance.

Finally, it should be noted that since the MCTN is designed for both bimodal and trimodal scenarios, its performance under unimodal conditions ($\{a\}, \{v\}, \{t\}$) may not be

Table 2
Quantitative results (%) for the case of the six available modality combinations. For example, "$\{t\}$" indicates that the text modality is available while the audio and video modalities are not present. "Avg" is the average performance of the six possible absent-modality combinations as shown in Table 1. "†" indicates results from [29]. **Bold**: best result. Underline: second best result. Higher values are better for all experimental results.

Datasets	Method	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{v,t\}$		$\{a,t\}$		Avg	
		ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
MOSI	MCTN	45.05	28.81	54.08	50.87	79.30	78.14	53.10	53.20	77.11	77.16	79.45	79.48	64.68	61.28
	TFR-Net	58.89	58.80	48.69	40.16	80.90	80.90	60.06	60.16	**81.63**	**81.65**	80.32	80.23	68.42	66.98
	MMIN	53.06	48.78	53.64	52.13	80.32	79.26	54.08	50.79	78.86	78.91	75.66	75.65	65.94	64.25
	If-MMIN	61.66	60.56	54.32	52.35	79.30	79.19	57.29	47.83	78.86	78.92	77.41	77.46	68.17	66.05
	MPMM†	57.26	59.35	58.63	59.12	79.81	80.10	60.54	61.33	80.74	80.93	79.89	79.84	69.48	70.11
	MPLMM†	62.71	63.65	63.12	63.74	80.12	80.31	65.02	65.41	81.12	81.19	80.76	81.09	72.14	72.57
	Ours	**72.16**	**72.22**	**70.41**	**70.47**	**81.34**	**81.10**	**73.03**	**73.02**	81.05	80.78	**81.63**	**81.39**	**76.60**	**76.50**
SIMS	MCTN	57.74	42.27	65.21	58.54	76.69	76.53	67.02	65.59	74.56	74.69	74.47	74.59	69.28	65.37
	TFR-Net	68.71	57.63	69.37	56.82	68.93	64.20	68.49	58.54	69.58	57.72	68.93	66.56	69.00	60.25
	MMIN	69.58	62.98	68.93	57.00	77.68	77.39	66.74	63.34	75.05	75.24	77.24	76.32	72.54	68.71
	If-MMIN	70.46	59.30	67.61	59.25	77.90	75.48	67.83	62.52	75.49	75.49	76.59	75.03	72.65	67.85
	MPMM†	64.98	76.41	65.40	77.92	78.56	78.56	64.01	73.47	77.51	77.47	77.11	77.20	71.26	76.85
	MPLMM†	65.93	**77.10**	66.02	**78.86**	79.75	78.74	65.28	74.02	77.97	77.95	77.45	77.84	72.07	**77.42**
	Ours	**72.65**	73.67	**73.52**	71.23	**80.74**	**79.78**	**73.30**	**74.24**	**80.96**	**80.16**	**80.09**	**80.11**	**76.88**	76.53
MOSEI	MCTN	67.29	60.88	70.96	59.16	80.15	80.69	71.00	58.98	80.60	81.02	81.05	81.44	75.18	70.36
	TFR-Net	71.02	58.99	63.96	63.51	80.45	80.92	58.08	60.02	74.09	75.28	81.28	81.62	71.51	70.06
	MMIN	70.77	59.22	56.90	58.86	79.93	80.61	63.45	64.81	79.24	79.91	75.98	77.06	71.05	70.08
	If-MMIN	70.27	60.22	71.17	59.77	80.27	80.83	53.36	54.91	77.57	78.33	79.39	80.05	72.00	69.02
	MPMM†	66.94	68.74	67.21	69.27	78.21	78.30	68.11	69.79	79.63	79.71	79.41	79.47	73.25	74.17
	MPLMM†	67.33	68.71	67.29	69.40	79.12	79.17	68.21	69.91	80.11	80.13	80.45	80.43	73.75	74.98
	Ours	**78.64**	**79.09**	**79.63**	**78.50**	**81.18**	**81.06**	**80.75**	**79.96**	**83.69**	**83.49**	**83.07**	**83.11**	**81.16**	**80.87**

Table 3
Quantitative results (%) for the full-modality availability case. Parenthetically, the results on MPMM and MPLMM is **not** reported in [29]. **Bold**: best result. Underline: second best result. Higher values are better for all experimental results.

Model	**MOSI**		**SIMS**		MOSEI	
	ACC2	F1	ACC2	F1	ACC2	F1
MCTN	78.13	78.14	76.89	76.71	79.82	80.31
TFR-Net	80.17	80.23	65.21	62.90	73.86	75.07
MMIN	81.20	80.98	74.84	75.35	81.95	82.23
If-MMIN	78.43	78.18	77.90	76.26	80.21	80.74
Ours	**81.49**	**81.43**	**79.65**	**78.52**	**83.22**	**83.26**

optimal.

4.5. Supplementary Experiments

We propose this model to cope with the diversity of absent modalities combinations in different scenarios, so it also includes scenarios where all three modalities are available. To further validate the generalisation ability of the model, we also conducted experiments on sentiment analysis under full modality conditions. As shown in Table 3, our model exhibits high and stable performance across different datasets. This demonstrates that our model can flexibly adapt to the needs of scenarios with different available modalities.

4.6. Ablation Experiments

We conducted ablation experiments to evaluate the roles of the Bilateral Auto-Encoder of Bilateral Imagination Module and the Pre-trained Encoders of Modality Encoder Network within the BiMIN. Notably, in the experiments where the core and peripheral branches were absent (Bilateral Auto-Encoders were not utilized), Auto-Encoders were used as substitutes for Bilateral Auto-Encoders. For the experiments where the audio and video modalities did not utilize Pre-trained Encoders, we employed the same feature extraction approach as used in [38] and the same encoders as used in [13].

Pre-trained Encoders. As shown in 4, the experimental results clearly demonstrate the performance of Pre-trained Encoders in handling scenarios where the textual modality is absent ($\{a\}$, $\{v\}$, $\{a,v\}$). A noticeable trend emerged from the comparison: introducing Pre-trained Encoders enhances model performance under these conditions. It is particularly noteworthy that both Experiment 1 and Experiment 2 utilized Auto-Encoders as the foundational structure for the Modality Imagination Network. Under the same conditions where the textual modality is absent, the results of Experiment 2 significantly surpassed those of Experiment 1. This strongly demonstrates that the advantage of performance enhancement provided by Pre-trained Encoders is not limited to specific models but is also applicable across different model architectures. Regarding the extent of performance improvement, we observed that in the presence of only the acoustic modality available ($\{a\}$), accuracy improved

Table 4
Quantitative results (%) for the case of the six absent-modality combinations for the ablation experiments on the dataset MOSEI with respect to the Pre-trained Encoder **PE**, the Core Branch **CB** and the Peripheral Branch **PB** of the Bilateral Auto-Encoder. A "$\surd$" indicates that the module was used. For example, "$\{t\}$" indicates that only the textual modality is available, while the audio and video modalities are absent. "Avg" is the average performance of the six possible absent-modality combinations as shown in Table 1. Higher values are approximately better for all the experimental results. **Bold**: best results. Underline: second best results.

No.	PE	CB	PB	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{v,t\}$		$\{a,t\}$		Avg	
				ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
1				70.77	59.22	56.90	58.86	79.93	80.61	63.45	64.81	79.24	79.97	75.98	77.06	71.05	70.09
2	$\surd$			75.66	76.04	72.72	73.79	77.85	78.64	75.25	75.48	78.28	75.99	80.15	80.67	76.65	76.77
3		$\surd$		67.89	66.41	55.89	57.69	79.93	80.49	63.75	64.91	80.70	81.19	80.30	80.84	71.41	71.92
4			$\surd$	68.92	63.08	55.18	57.09	77.83	78.69	59.80	61.66	79.74	80.40	79.91	80.36	70.23	70.21
5	$\surd$	$\surd$		74.33	75.13	69.11	68.40	79.54	79.58	76.20	75.84	79.85	78.68	80.08	80.59	76.52	76.37
6	$\surd$		$\surd$	73.38	74.05	72.05	72.89	79.37	78.94	75.04	74.82	74.50	75.66	77.57	78.44	75.32	75.80
7		$\surd$	$\surd$	70.04	65.00	63.13	64.00	80.88	81.35	69.74	66.37	82.49	82.43	81.43	81.89	74.62	73.46
8	$\surd$	$\surd$	$\surd$	**78.64**	**79.09**	**79.63**	**78.50**	**81.18**	**81.06**	**80.75**	**79.96**	**83.69**	**83.49**	**83.07**	**83.11**	**81.16**	**80.87**

by 5% to 8%; in the presence of only the visual modality available ($\{v\}$), accuracy increased by 14% to 16%; and in the combined presence of only both acoustic and visual modalities available ($\{a,v\}$), accuracy gains ranged from 11% to 15%.

Bilateral Auto-Encoder. As shown in Table 4, Experiments 1 and 2 utilized Auto-Encoders as the core architecture of the Modality Imagination Network, while Experiments 7 and 8 employed complete Bilateral Auto-Encoders as the backbone. Despite using the same Pre-trained Encoders, we observed that the average performance in Experiment 7 exceeded that of Experiment 1 by 3%, and Experiment 8 outperformed Experiment 2 by 3% as well. To further explore the necessity of the dual-branch structure in bilateral Auto-Encoders, we compared Experiments 3 and 4 with Experiment 7, as well as Experiments 5 and 6 with Experiment 8. The results indicate that experiments equipped with complete Bilateral Auto-Encoders (i.e., Experiments 7 and 8) demonstrated an average performance improvement of 3% to 5% compared to those with only a single branch (i.e., Experiments 3, 4, 5, and 6). This finding underscores the significant advantage of the dual-branch structure in enhancing performance.

In all experimental configurations, Experiment 8 is particularly noteworthy as it combines both Pre-trained Encoders and Bilateral Auto-Encoders. When evaluated across six different modality availability scenarios, Experiment 8 consistently delivered the best performance, with enhancements that were both balanced and stable. By examining the second-best results, we observe that models incorporating pre-trained encoders exhibit performance gains when the textual modality is absent, while those using bilateral autoencoders demonstrate improvements when the textual modality is present.

To further validate the performance advantages of the Bilateral Auto-Encoder, we conducted a controlled experiment under identical conditions. Specifically, we replaced the Auto-Encoders in the MMIN and If-MMIN models with Bilateral Auto-Encoders and compared the experimental results, as summarized in Table 5. The data clearly reveal a trend: when models are equipped with Bilateral Auto-Encoders, significant improvements are achieved across multiple performance metrics. This enhancement is particularly pronounced in multi-classification tasks, with the If-MMIN model showing especially notable performance gains. These findings not only confirm the Bilateral Auto-Encoders' capability to capture data complexity but also highlight its potential to significantly enhance overall model performance.

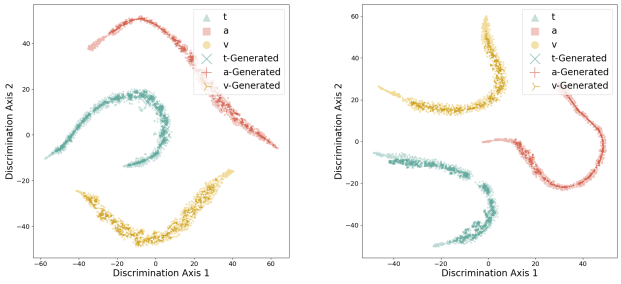
Table 5
On ablation experiments with bilateral encoders. We report the average performance for the six available modalities combinations. Where "**MMIN**" and "**If-MMIN**" denotes the original backbone network using the Auto-Encoder; "**MMIN(BiAE)**" and "**If-MMIN(BiAE)**" denotes that the Auto-Encoders was replaced with Bilateral Auto-Encoders. "↑" indicates that higher is better and "↓" indicates that lower is better.

model	ACC2↑	F1↑	ACC5↑	ACC7↑	MAE↓
MMIN	71.05	70.08	45.43	44.66	0.723
MMIN(BiAE)	74.37	71.15	47.81	47.14	0.699
If-MMIN	72.01	69.02	41.57	40.47	0.820
If-MMIN(BiAE)	74.51	73.59	47.53	46.80	0.684

4.7. Capability Analysis

The core advantage of the Bilateral Multimodal Imagination Network (BiMIN) lies in its unique Bilateral Imagination Module, which is specifically designed to generate representations for absent modalities. To empirically validate the effectiveness of this module in generating representations for absent modalities, we conducted a series of meticulously designed visualization experiments. As illustrated in Figure 4, we utilized t-SNE [40] to compare the visual distributions of the ground-truth multimodal embeddings (as shown by h in Figure 2) with the multimodal embeddings generated by Bi-MMIN (denoted as $\hat{h}$ in Figure

2). For the MOSEI test set, we randomly selected 512 sets of multimodal data for analysis, while for the SIMS test set, we included all 457 sets of multimodal data for comprehensive testing. The results demonstrate a high degree of similarity between the distributions of the generated multimodal embeddings and the ground-truth multimodal embeddings, across both the English MOSEI dataset and the Chinese SIMS dataset. Notably, the generation quality for all three modalities—text, audio, and video—was consistently robust. This finding confirms that BiMIN is capable of generating representations for absent modalities based on the available ones, providing an effective solution for addressing the common issue of modality uncertainty in real-world scenarios.



(a) English Dataset **CMU-MOSEI** (b) Chinese Dataset **CH-SIMS**

Fig. 4. Visualisation of grand-truth multimodal embeddings and generated multimodal embeddings. For example, "**t**" denotes a grand-truth embeddings of textual modality, and "**t-Generated**" denotes embeddings of textual modality generated based on the visual modality and the acoustic modality.

5. Conclusion

In this paper, we introduce a novel sentiment analysis model designed to address modality uncertainty—the Bilateral Modality Imagination Network (BiMIN). This model comprises two key components: a Modality Encoder Network based on pre-trained feature extractor, and a Bilateral Imagination Module based on the Bilateral Auto-Encoder. The Modality Encoder Network is adept at extracting sentiment from available modalities, thereby enhancing overall performance. Meanwhile, the Bilateral Imagination Module improves the robustness of joint multimodal representations and the accuracy of predictions. Experimental results on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets demonstrate that BiMIN consistently outperforms baseline models across various absent-modality combinations. In future work, we aim to further refine the Modality Encoder Network and the Bilateral Imagination Module.

References

[1] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[2] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the conference. Association for computational linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, p. 6558.

[3] S. Liu, W. Quan, Y. Liu, D.-M. Yan, Bi-directional modality fusion network for audio-visual event localization, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 4868–4872.

[4] Y. Lin, P. Ji, X. Chen, Z. He, Lifelong text-audio sentiment analysis learning, Neural Networks 162 (2023) 162–174.

[5] C. Li, J. Wang, H. Wang, M. Zhao, W. Li, X. Deng, Visual-texual emotion analysis with deep coupled video and danmu neural networks, IEEE Transactions on Multimedia 22 (6) (2019) 1634–1646.

[6] D. Tiwari, B. Nagpal, Keaht: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis, New Generation Computing 40 (4) (2022) 1165–1202.

[7] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, Q. Zheng, Learning multi-scale features for speech emotion recognition with connection attention mechanism, Expert Systems with Applications 214 (2023) 118943.

[8] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, Z. Luo, Clip-aware expressive feature learning for video-based facial expression recognition, Information Sciences 598 (2022) 182–195.

[9] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1122–1131.

[10] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, Vol. 2020, NIH Public Access, 2020, p. 2359.

[11] J. Guo, J. Tang, W. Dai, Y. Ding, W. Kong, Dynamically adjust word representations using unaligned multimodal information, in: Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 3394–3402.

[12] L. Tran, X. Liu, J. Zhou, R. Jin, Missing modalities imputation via cascaded residual autoencoder, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1405–1414.

[13] J. Zhao, R. Li, Q. Jin, Missing modality imagination network for emotion recognition with uncertain missing modalities, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 2608–2618.

[14] H. Zuo, R. Liu, J. Zhao, G. Gao, H. Li, Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[15] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] C. Zheng, J. Peng, L. Wang, L. Zhu, J. Guo, Z. Cai, Frame-level nonverbal feature enhancement based sentiment analysis, Expert Systems with Applications 258 (2024) 125148.

[17] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 10790–10797.

[18] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, L.-P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 7216–7223.

[19] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 6892–6899.

[20] Z. Yuan, W. Li, H. Xu, W. Yu, Transformer-based feature reconstruction network for robust multimodal sentiment analysis, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4400–4407.
[21] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).
[22] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, W. Kong, Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5301–5311.
[23] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
[24] D. P. Kingma, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[25] G. Pandey, A. Dukkipati, Variational methods for conditional multimodal deep learning, in: 2017 international joint conference on neural networks (IJCNN), IEEE, 2017, pp. 308–315.
[26] C. Du, C. Du, H. Wang, J. Li, W.-L. Zheng, B.-L. Lu, H. He, Semi-supervised deep generative modelling of incomplete multi-modality emotional data, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 108–116.
[27] J. Zeng, J. Zhou, T. Liu, Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 2924–2934.
[28] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, C.-Y. Lee, Multimodal prompting with missing modalities for visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14943–14952.
[29] Z. Guo, T. Jin, Z. Zhao, Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition, arXiv preprint arXiv:2407.05374 (2024).
[30] J. Zeng, T. Liu, J. Zhou, 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, 2022, pp. 1545–1554.
[31] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE signal processing letters 23 (10) (2016) 1499–1503.
[32] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: 2016 IEEE winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10.
[33] A. Zadeh, Y. C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit tadas baltrušaitis, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2018.
[34] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, H. Ling, Expanding language-image pretrained models for general video recognition, in: European Conference on Computer Vision, Springer, 2022, pp. 1–18.
[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.
[36] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intelligent Systems 31 (6) (2016) 82–88.
[37] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
[38] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 3718–3727.
[39] D. P. Kingma, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[40] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11) (2008).