BiMIN: Bilateral Modality Imagination Network for Sentiment Analysis under Modalities Uncertainty

Xuanchao Lin[a], Junjie Peng[a,b,*], Yijie Jin[a], Jiahao Guo[a] and Zesu Cai[c]

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

ARTICLE INFO

Keywords:
Multimodal Sentiment Analysis
Modality Reconstruction
Unified Model

ABSTRACT

Recent studies have shown that Multimodal Sentiment Analysis, which utilizes textual, acoustic, and visual modalities, improves the accuracy of Sentiment Analysis. However, in practical scenarios, it is frequently difficult to obtain all three modalities simultaneously. Depending on the specific circumstances, only one or two modalities might be available, with the remaining ones being absent. Under such modalities uncertainty, models that rely on a particular set of modalities for sentiment analysis can suffer from severe ineffectiveness. The absence of one or more modalities can result in a substantial decline in the model's predictive capabilities. This highlights the need for more robust and adaptable models that can maintain accuracy even when some modalities are absent. Given the potential absence of one or more modalities, much of the prior research has focused on reconstructing the absent modalities using the available ones, yielding some promising, yet limited results. Building on this foundation and aiming to further address the limitations of current approaches, we propose the Bilateral Modality Imagination Network (BiMIN), a unified model designed to handle various scenarios of modality absence. BiMIN incorporates two key components: 1) Modality Encoder Network utilizing pre-trained models to mine deep sentiment information from available modalities; 2) Bilateral Imagination Module to improve the performance of reconstruction of absent modalities. Comprehensive experiments on three 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 gather information from their surroundings through various senses to form an understanding of the world. The environment is rich with different types of information, such as text, sound, and images, which serve as the primary media for information transmission. In this context, sentiment analysis is primarily used to assess a person's emotional tendencies through textual, acoustic, or visual modalities, or a combination of two or more.

In real life, the types of modality available to people in different situations are not certain. For instance, during a phone conversation, only voice and the corresponding text information can be accessed. When interacting with someone who speaks a different language, only facial expressions, movements, and voice are perceivable, while the spoken text remains unintelligible. In the case of a patient who is temporarily aphasic but can still communicate by writing or typing, only text and visual information are accessible. Situations like the ones described above are common in reality, where only two or even a single modality is available during communication, as shown in the Table 1.

However, most general models for sentiment analysis are based on ideal, certain scenarios. For example, Tensor Fusion Network (TFN) (Zadeh, Chen, Poria, Cambria and Morency, 2017) and Cross-Modal Hierarchical Fusion Model (CMHFM) (Wang, Peng, Zheng, Zhao and Zhu, 2024) are developed for trimodal. Models like Bi-directional Modality Fusion Network (BMFN) (Liu, Quan, Liu and Yan, 2022a), Lifelong Text-Audio Sentiment Analysis (LTASA) (Lin, Ji, Chen and He, 2023), and Deep Coupled Video and Danmu Neural Networks (DCVDN) (Li, Wang, Wang, Zhao, Li and Deng, 2019) focus on bimodal, addressing audio-video, text-audio, and video-text modalities, respectively. Additionally, Tiwari and Nagpal (2022) address unimodal textual analysis, Chen, Li, Liu, Wang, Wang and Zheng (2023) focus on unimodal acoustic analysis, and Liu, Feng, Yuan, Zhou, Wang, Qin and Luo (2022b)

*Corresponding author
linxuanchao@shu.edu.cn (X. Lin); jjie.peng@shu.edu.cn (J. Peng); jyj2431567@shu.edu.cn (Y. Jin); 13681664265@shu.edu.cn (J. Guo); caizesu@hit.edu.cn (Z. Cai)

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)

explore unimodal visual analysis. Given the diverse real-world scenarios previously discussed, it is important to acknowledge that sentiment analysis models tailored for specific modalities can become highly ineffective when some modalities are absent. This highlights the need for a thorough redesign of network architecture. For instance, a multimodal sentiment analysis model optimized for trimodal input—designed to handle textual, acoustic, and visual modalities—may encounter significant challenges in bimodal or unimodal scenarios. The absence of one or more modalities can result in a notable decline in performance, as the model's predictive capabilities heavily rely on the modality availability in certain scenarios it is initially designed for.

When conducting sentiment analysis in scenarios where the availability of modalities is uncertain, two primary strategies are typically employed:

1. **Maximizing Utilization of Available Modalities**: Given the absence of modalities, it is crucial to extract the maximum amount of sentiment-related information from the available modalities. This requires a tailored strategy to effectively mine sentiment information from the diverse modalities present.
2. **Imagination or Generation of Absent Modalities**: Given the varying absence of modalities, a unified strategy is needed to handle these situations effectively. Generating synthetic data for the absent modalities or imagining their characteristics from the available data can help standardize this process, thereby improving sentiment analysis.

In essence, to address the challenges posed by the varying absence of modalities, it's necessary to develop more adaptable and robust models for sentiment analysis. These models must not only process the available data but also compensate for the absence of certain modalities through innovative data synthesis or inference techniques. To achieve this goal, we propose **Bilateral Modality Imagination Network (BiMIN)**, a unified model designed to handle sentiment analysis under modality uncertainty. In addition to managing all six possible combinations of absent modalities, BiMIN can effectively perform sentiment analysis when all three modalities are available. BiMIN delves deeper into the sentiment information of each available modality and enhances the robustness of joint multimodal representations, leading to significant improvements in sentiment analysis performance. BiMIN adopts two core mechanisms:

1. Leverages pre-trained models to extract and encode feature information from the available modalities, optimizing their utility.
2. Employs Bilateral Auto-Encoder we design to imagine or generate absent modalities, further enhancing interaction and integration between modalities.

Maximizing Utilization of Available Modalities. Studies (Hazarika, Zimmermann and Poria, 2020; Rahman, Hasan, Lee, Zadeh, Mao, Morency and Hoque, 2020; Guo, Tang, Dai, Ding and Kong, 2022) have shown that the textual modality is the most dominant in sentiment analysis. Our experiments (as shown in subsection 4.4) also verify a sharp decline in performance when textual modality is absent. This is due to two key factors: textual modality explicitly conveys sentiment through specific words (e.g. "$like$," "$hate$"), and pre-trained models like BERT (Devlin, Chang, Lee and Toutanova, 2019) can capture advanced semantic features, greatly boosting sentiment analysis performance. In contrast, non-textual modalities often struggle due to the complexity of their sub-features, which hinders effective integration and leads to bottlenecks. By applying pre-trained models to non-textual modalities, BiMIN aims to overcome these challenges. As a result, while many models falter without text, BiMIN maintains strong performance even when textual modality is absent.

Imagination or Generation of Absent Modalities. Many studies (Tran, Liu, Zhou and Jin, 2017; Zhao, Li and Jin, 2021; Zuo, Liu, Zhao, Gao and Li, 2023) have shown that the architecture of Cascaded Residual Auto-Encoders can generate absent modalities while creating joint multimodal representations for sentiment analysis. Therefore, we adopt this approach to tackle the challenges arising from modalities uncertainty. However, the single-branch design of traditional Auto-Encoder often leads to the loss of fine-grained details, creating a performance bottleneck in sentiment prediction. Our designed Bilateral Auto-Encoder addresses this issue by incorporating two independent yet collaborative branches: the Core Branch preserves coarse-grained core information, and the Detail Branch preserves fine-grained detail information. This dual-branch architecture effectively reduces detail loss and enhances sentiment analysis performance compared to single-branch designs.

Experimental results on three public datasets for Multimodal Sentiment Analysis indicate that the model consistently outperforms benchmark models across all combinations of available modalities.

The primary contributions of this work are:"

- We design a Bilateral Modality Imagination Network (BiMIN) aiming to address sentiment analysis under modality uncertainty.
- BiMIN leverages pre-trained models to extract features from various modalities, thereby enhancing the usability and expressiveness of available modalities.
- The Bilateral Auto-Encoder we designed is capable of simultaneously capturing and preserving both coarse-grained core information and fine-grained detail information of the modalities, forming a Bilateral Imagination Module that enhances the robustness of joint multimodal representations.
- Comprehensive experimental results on three major multimodal sentiment analysis datasets show that our model greatly enhances sentiment analysis performance under modality uncertainty.

2. Related Work

2.1. Sentiment Analysis under Modalities Certainty

Sentiment Analysis is a computational technique used to automatically detect and interpret human emotions, attitudes, and sentiments conveyed through various modalities. Initially, this analysis focused on a single modality, such as textual modality (Tiwari and Nagpal, 2022), acoustic modality (Chen et al., 2023), or visual modality (Liu et al., 2022b). However, relying on just one modality often results in limitations in sentiment prediction.

To improve accuracy, multimodal sentiment analysis has emerged. It combines multiple modalities—such as textual modality, acoustic modality, or visual modality—to enhance sentiment prediction through mutual reinforcement. Despite its advantages, multimodal sentiment analysis faces a key challenge: how to effectively integrate information from various modalities so they can complement one another and lead to more accurate sentiment predictions. Additionally, variations in real-world scenarios result in differences in the available modalities, necessitating tailored models and frameworks for each specific combination of modalities.

Models for Bimodal Modality Analysis. Liu et al. (2022a) focus on **audio-video fusion**, presenting a mechanism that enhances the representation of acoustic and visual features through bidirectional fusion, refining these features via a feedforward-backward attention module. Lin et al. (2023) center their work on **audio-text fusion**, designing a complementary-aware subspace to explore the nonlinear complementary knowledge between acoustic and textual modalities. In **video-text fusion**, Li et al. (2019) integrate video content with real-time text, utilizing deep learning to synchronously extract and fuse visual and textual features, employing a Deep Canonically Correlated Auto-Encoder for multi-view learning to enable precise analysis.

Models for Trimodal Modality Analysis. TFN (Zadeh et al., 2017) models relationships between modalities through the Cartesian product. Tsai, Bai, Liang, Kolter, Morency and Salakhutdinov (2019) tackle modality misalignment and heterogeneous information fusion by aligning modalities through a cross-modal attention mechanism. The model proposed by Hazarika et al. (2020) learns modality invariance as well as modality specificity respectively by projecting each modality into two distinct subspaces. Yu, Xu, Yuan and Wu (2021) implement a self-supervised approach to create unimodal labels while simultaneously training on both multimodal and unimodal tasks. The model proposed by Zhao, Peng, Huang, Wang, Zhang and Cai (2023) bridges semantic gaps, enhances low-density features with convolutional aggregation, and improves long-term sequence modeling through dynamic routing. CMHFN (Wang et al., 2024) leverages interactions across unimodal, bimodal, and trimodal inputs to strengthen the model's resilience.

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

2.2. Sentiment Analysis under Modalities Uncertainty

Due to the uncertainty of scenarios in reality, the absence of modalities can cause sentiment analysis models designed for certain modality scenarios to fail. To address this issue, previous research has explored effective multimodal sentiment analysis using various strategies, primarily including attention-based, Auto-Encoder-based, variational Auto-Encoder-based, and optimization strategies derived from these approaches.

Models Based on Attention. Multimodal Cyclic Translation Network (MCTN) (Pham, Liang, Manzini, Morency and Póczos, 2019) employs a sequence-to-sequence model for cyclic translation between modalities, offering robustness to noise and missing data while learning strong joint representations. Transformer-based Feature Reconstruction Network (TFR-Net) (Yuan, Li, Xu and Yu, 2021) handles randomly missing data in unaligned modality sequences using internal and cross-modal attention mechanisms. Tang, Li, Jin, Cichocki, Zhao and Kong (2021) enhance dual multimodal fusion embedding with a hierarchical structure of multiple bi-directional translations, outperforming MCTN in this aspect. Despite their effectiveness in capturing inter-modality relationships, attention-based models are computationally intensive, lack real-time performance, and struggle when a modality is significantly missing or absent.

Models Based on Auto-Encoder. Auto-Encoder-based models learn latent representations to generate data with distributions similar to the observed inputs, making them highly applicable to multimodal sentiment analysis under modalities uncertainty. Tran et al. (2017) address absent modalities by utilizing a cascading architecture of residual Auto-Encoders to learn relationships across modalities. Missing Modality Imagination Network (MMIN) (Zhao et al., 2021) extends this by incorporating Cycle Consistency Learning (Zhu, Park, Isola and Efros, 2017), enhancing its ability to predict absent modality representations and operate under various absent conditions. Invariant Features for a Missing Modality Imagination Network (If-MMIN) (Zuo et al., 2023) further improves upon MMIN with central moment discrepancy distance constraints to learn modality-invariant features, tackling real-world scenarios of modalities uncertainty. Although these Auto-Encoder-based models perform well under modality uncertainty, they still face limitations due to the inherent single-branch design of Auto-Encoders, leading to performance bottlenecks.

Models Based on Variational Auto-Encoder. Research based on Variational Auto-Encoder (VAE) enhances Auto-Encoder capabilities by integrating variational inference, improving data generation and feature extraction. 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) (Pandey and Dukkipati, 2017) 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) (Du, Du, Wang, Li, Zheng, Lu and He, 2018) models the statistical relationships in multimodal sentiment data through multiple modality-specific generative networks and a shared latent space. By applying a Gaussian mixture assumption to the posterior approximation of the shared latent variables, MVAE learns joint deep representations across modalities and assesses the importance of each modality. Despite their strengths in generating samples and inferring latent variables, VAE-based models face challenges such as high computational complexity, training difficulty, and limitations from the Gaussian assumption on latent space distributions.

Optimization Strategies. Recent works have introduced optimization strategies based on the aforementioned approaches. Ensemble-based Missing Modality Reconstruction (EMMR) (Zeng, Zhou and Liu, 2022b) combines Auto-Encoder and attention-based methods to enhance decision-making capabilities through integration techniques. However, it faces limitations, including the single-branch design constraints of Auto-Encoders and the computational intensity and lack of real-time performance associated with attention mechanisms. Multimodal Prompting with Missing Modalities (MPMM) (Lee, Tsai, Chiu and Lee, 2023) and Multimodal Prompt Learning with Missing Modalities (MPLMM) (Guo, Jin and Zhao, 2024) integrate prompt information into the Transformer architecture to optimize performance. Although the use of prompt information is intended to make the model more lightweight, both still require pre-training the Transformer for different datasets, resulting in high computational complexity despite the optimization efforts. Tag-Assisted Transformer Encoder (TATE) (Zeng, Liu and Zhou, 2022a) introduces a label encoding module to tag missing modalities, which guides the Transformer network to prioritize the missing data. Additionally, it employs a

novel spatial projection method to align shared feature vectors across modalities. The limitations of Transformer-based approaches will not be further elaborated here.

Compared to previous work, we have designed the Bilateral Auto-Encoder within the Auto-Encoder framework and developed the Bilateral Modality Imagination Network (BiMIN). This model achieves a balance between accuracy and efficiency, effectively addressing the challenges of sentiment analysis in scenarios with uncertain modalities. Notably, BiMIN significantly improves performance, particularly in cases where the textual modality is absent, demonstrating substantial enhancements under these conditions.

3. Method

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

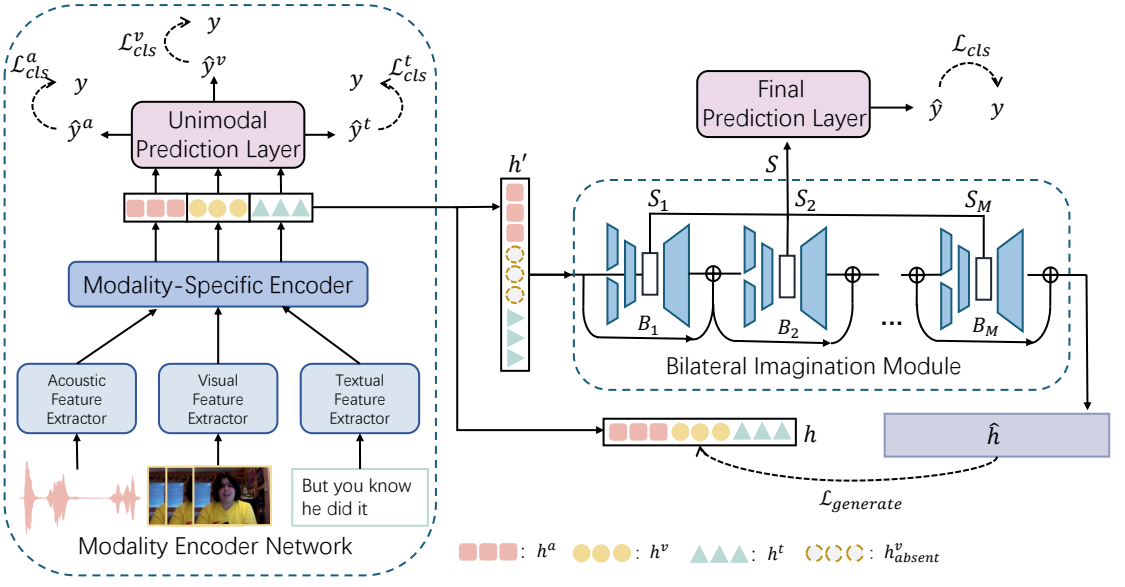


Fig. 1: Illustration of the Bilateral Modality Imagination (BiMIN) framework.

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 proposed model 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 Fig. 1, 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 illustrated in Fig. 1, comprises three primary 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-Specific Encoder to obtain the full-modality embeddings h.

Bilateral Imagination Module: Given the initial embedding of the available modalities h', the Bilateral Imagination Module, consisting of M Bilateral Auto-Encoders, generates imagined multimodal embedding $\hat{h}$. In this process,

hidden states S_i from each Bilateral Auto-Encoder B_i are collected to form a joint multimodal representation S, where $i \in \{1, 2, ..., M\}$.

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

The subsequent subsections will offer a thorough explanation of each component.

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 use a pre-trained BERT (Devlin et al., 2019) language model as our textual feature extractor. BERT, a Transformer model trained on a large corpus in a self-supervised fashion, captures rich sentiment information through learned inner representations of text. For a given sentence $x^t = \{w_1, w_2, ..., w_n\}$ consisting of n words, we append the special tokens [CLS] and [SEP], and feed this sequence into BERT to extract the textual modality features F^t.

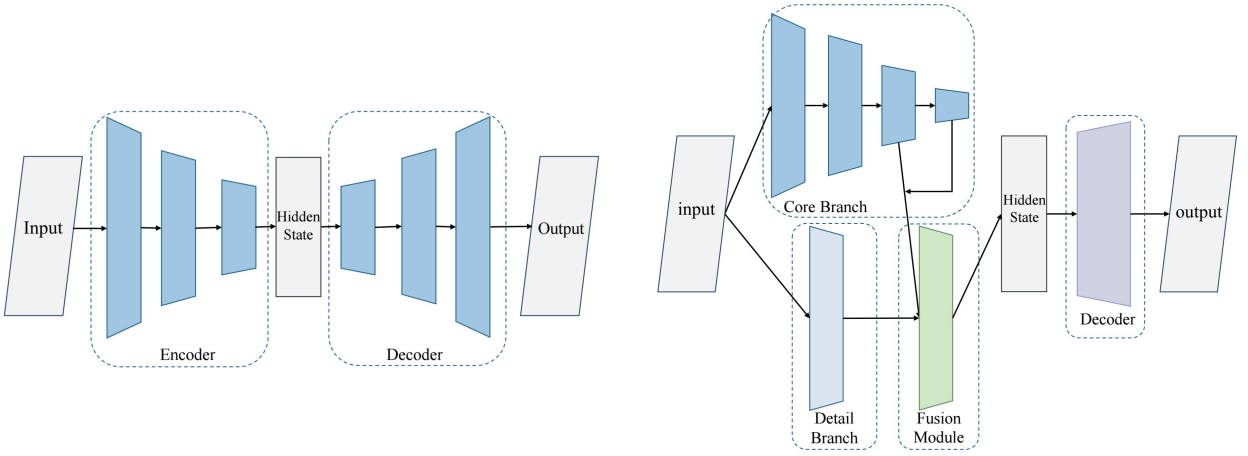
Visual Feature Extractor. In the visual modality, previous works commonly employed Multi-task Cascaded Convolutional Networks (MTCNN) (Zhang, Zhang, Li and Qiao, 2016) or OpenFace (Baltrušaitis, Robinson and Morency, 2016; Zadeh, Lim and Morency, 2018a) to capture facial expression features. These features include 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, in previous approaches, these sub-features are often directly concatenated to form the visual feature, which limits the performance of sentiment analysis tasks. To address this issue, we draw inspiration from textual feature extraction methods and employ the pre-trained visual model X-CLIP (Ni, Peng, Chen, Zhang, Meng, Fu, Xiang and Ling, 2022) as a visual feature extractor to capture the raw features from videos. Specifically, we uniformly sample 8 frames from the original video x^v and convert them into the required tensor format using the preprocessing functions provided by X-CLIP. The processed data is then fed into X-CLIP to extract the visual features F^v. The features extracted by this method are treated as a unified whole, rather than as concatenated heterogeneous sub-features, allowing the subsequent network to better utilize the sentiment information from the visual modality.

Acoustic Feature Extractor. In previous research, tools like OpenSMILE (Eyben, Wöllmer and Schuller, 2010), COVAREP (Degottex, Kane, Drugman, Raitio and Scherer, 2014), and LibROSA (McFee, Raffel, Liang, Ellis, McVicar, Battenberg and Nieto, 2015) are commonly used to extract acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, intensity, and peak frequency. While these sub-features are structurally distinct, they are interrelated and together capture the sentiment information in speech. However, as with the visual modality, these sub-features are often directly concatenated as input to the model, which makes it challenging for the model to effectively recognize and learn from the heterogeneous data, leading to performance bottlenecks. We employ Whisper (Radford, Kim, Xu, Brockman, McLeavey and Sutskever, 2023), a large pre-trained model designed for automatic speech recognition and translation to address this issue. Specifically, we convert the raw audio data x^a into a spectrogram using a Mel filter and apply a logarithmic scale to enhance high-frequency details. The transformed data is then padded or trimmed to match Whisper's required input size. Finally, the processed data is fed into Whisper's encoder to extract the acoustic modality features F^a. These features are treated as a unified whole rather than as concatenated heterogeneous sub-features, enabling the subsequent network to better utilize the sentiment information from the acoustic modality.

Modality-Specific Encoder: We adopt a stagewise training approach to process the extracted textual features F^t, visual features F^v, and acoustic features F^a. Specifically, we use recurrent neural networks (RNNs) to capture the temporal relationships among these features and unify their feature dimensions, resulting in textual embedding h^t, visual embedding h^v, and acoustic embedding h^a. These embeddings are concatenated to form the full-modality embedding $h = (h^t, h^v, h^a)$. For scenarios with varying numbers of available modalities, we represent an absent-modality embedding h^m as a zero vector h^m_{absent} 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 of available modalities for the Bilateral Imagination Module would be $h' = (h^t, h^v_{absent}, h^a)$. In the subsequent subsection 4.2, we use a unified input format as (x^t, x^v_{absent}, x^a). Since our approach involves stagewise training, there is no substantive difference between the two formats for the Bilateral Imagination Network.

3.4. Bilateral Imagination Module

Auto-Encoder is 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 state through a nonlinear mapping and then decodes it back through another mapping, as illustrated in Fig. 2a. The CRA network, illustrated in Fig. 3, builds on this by connecting multiple Auto-Encoders in a cascaded residual architecture. It extracts hidden states from each Auto-Encoder to form a joint multimodal representation for sentiment polarity prediction. This approach effectively imagines absent modalities and predicts sentiment polarity, which is why we have adopted the same strategy in the Bilateral Imagination Module. However, due to the single-branch structure of the original Auto-Encoder, its channel capacity is relatively low, inevitably leading to information compression. This causes the hidden state to retain only coarse-grained core information, losing fine-grained details. Although residual connections are added after each decoder to restore some of the original information in the subsequent Auto-Encoder, the joint multimodal representation formed by the hidden states still lacks fine-grained detail information. This limitation creates a bottleneck in final prediction performance, as the loss of fine-grained detail information negatively impacts the accuracy of emotion prediction, especially in fine-grained multi-class prediction tasks.



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

Fig. 2: Illustration of two different Auto-Encoder architectures. (a) shows the original Auto-Encoder, which consists of a single branch. (b) illustrates the Bilateral Auto-Encoder that we designed. This architecture has two branches, the Core Branch for coarse-grained core semantics and the Detail Branch for fine-grained details. the core branch has narrow channels and deep layers, while The detail branch has wide channels and shallow layers.

To address this issue, we present the **Bilateral Auto-Encoder**, as illustrated in Fig. 2b. The Bilateral Auto-Encoder includes a core branch and a detail branch, designed to capture both coarse-grained core information and fine-grained detail information. By doing so, it mitigates performance loss caused by the loss of detailed information.

The **Detail 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 Detail Branch requires substantial channel capacity to retain the fine-grained detailed sentiment 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 Detail 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 Detail 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-modal attention mechanisms. After considering both accuracy and

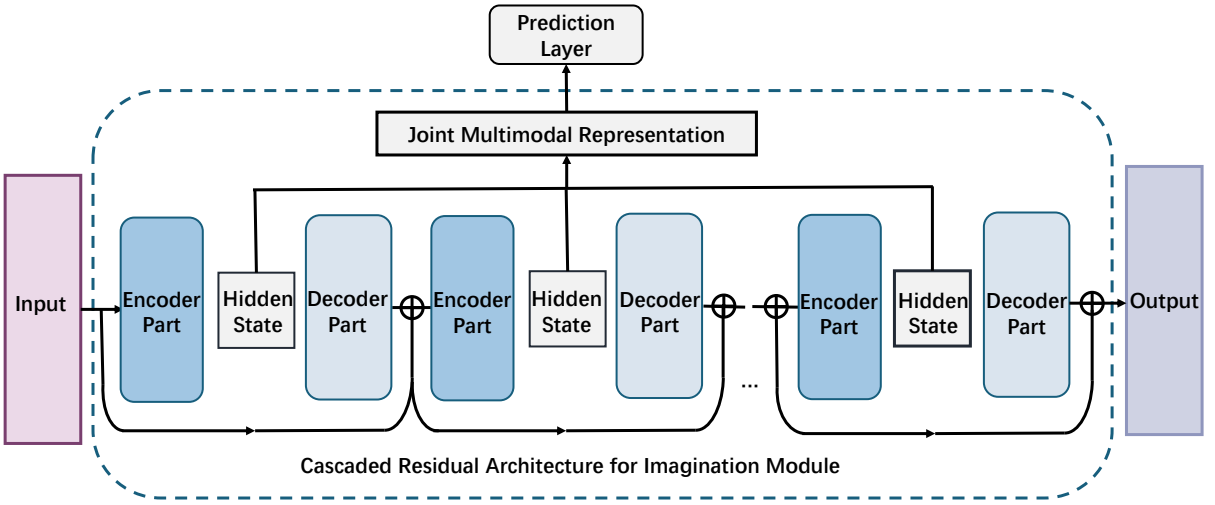


Fig. 3: Illustration of the Cascaded Residual Architecture for Imagination Module. The Encoder Part and Decoder Part as a unified entity, which can either be an **Original Auto-Encoder** or a **Bilateral Auto-Encoder** designed by us.

efficiency, we choose 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 consider both accuracy and efficiency, discarding the typical three-layer linear design used in the original Auto-Encoder (as shown in Fig. 2a) and opting instead for a single-layer linear decoder.

As illustrated in Fig. 1, Bilateral Imagination Module consists of M Bilateral Auto-Encoders. These Bilateral Auto-Encoders are connected in a cascaded residual architecture, as shown in Fig. 3. The structure of each Bilateral Auto-Encoder B_i is depicted in Fig. 2b, where the core branch is denoted as cb_i, the Detail Branch as db_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:

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

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 Fig. 1), 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. Sentiment Prediction Layer

As illustrated in Fig. 1, it is evident that the Modality Encoder Network contains a Unimodal Prediction Layer. Additionally, a Final Prediction Layer is connected outside the Bilateral Imagination Module.

The **Unimodal Prediction Layer** corresponds to a sentiment prediction layer PL^m for each modality. The embedding of each modality h^m ,as mentioned in subsection 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-Specific Encoder.

The **Final Prediction Layer** is a sentiment prediction layer PL, used for the final overall sentiment polarity prediction:

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

Where $PL(\cdot)$ denotes the function of the 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 $\mathcal{L}^m_{cls}$, which is defined as follows:

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

Where $H(y, \hat{y}^m)$ represents the cross-entropy between distributions ground-truth label y and modality-specific prediction $\hat{y}^m$, and $m \in \{t, v, a\}$.

In the Bilateral Imagination Network, two 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':

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

Where, $H(y, \hat{y})$ represents the cross-entropy between distributions ground-truth label y and final sentiment prediction $\hat{y}$, and $\|\cdot\|_2^2$ is the squared Frobenius norm. The overall loss $\mathcal{L}$ is calculated by adding these two functions together:

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

Where λ is a hyperparameter.

4. Experiments

4.1. Datasets and Evaluation Metrics

To replicate real-world conditions, we assessed our model using three well-established datasets for multimodal sentiment analysis: CMU-MOSI (Zadeh, Zellers, Pincus and Morency, 2016), CMU-MOSEI(Zadeh, Liang, Poria, Cambria and Morency, 2018b), and CH-SIMS (Yu, Xu, Meng, Zhu, Ma, Wu, Zou and Yang, 2020).

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 -3 (strongly negative) to +3 (strongly positive).

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 -1 (strongly negative) to +1 (strongly positive).

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 binary accuracy (Acc), F1 score (F1), 5-class accuracy (Acc5), 7-class accuracy (Acc7), and Mean Absolute Error (MAE).

4.2. Experimental Setup

Feature Extraction. As a baseline comparison, we extract frame-level raw features for each modality across all three datasets using the same method as (Yu et al., 2020).

Training Set 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 Fig. 1, the unified format of the available modality test samples would consist of the original features (x^t, x^v_{absent}, x^a). These six subsets, each representing a modality absence scenario, serve as the test sets.

Model Training Details. In all experiments, we employ the Adam optimizer (Kingma and Ba, 2015) with a batch size of 64. The model begins training with an initial learning rate of 1×10^{-3} using early stopping with patience of 8 epochs. To ensure reproducibility, we fix the random seed so that all models are trained on the same dataset. All models are constructed by employing the PyTorch deep learning framework and are run on a single Nvidia RTX 4060 Ti GPU.

4.3. Baselines

We contrast our proposed model with the following:

- **MCTN**: Multimodal Cyclic Translation Network (Pham et al., 2019) learns robust joint representations through modality translation to process missing information.
- **TFR-Net**: Transformer-based Feature Reconstruction Network (Yuan et al., 2021) uses Transform architecture to supplement missing information for emotion prediction.
- **MMIN**: Missing Modality Imagination Network (Zhao et al., 2021) learns robust joint multimodal representations that can infer the representation of any missing modalities based on the available modalities.
- **If-MMIN**: Invariant Features for a Missing Modality Imagination Network (Zuo et al., 2023) extracts invariant features between different modalities in order to learn robust joint multimodal representations, enabling the prediction of representations for any missing modalities based on the available ones.
- **MPMM**: Multimodal Prompting with Missing Modalities (Lee et al., 2023) uses deletion-aware cues to guide models in solving missing modalities problems.
- **MPLMM**: Multimodal Prompt Learning with Missing Modalities (Guo et al., 2024) uses a variety of cues to guide the generation of missing modalities features.

4.4. Main Results

The comparison of individual models under different absent-modality combinations is presented in Table 2. Analyzing the quantitative results across the three datasets, It is clear that our proposed BiMIN consistently exceeds the performance of the baselines across all datasets and all six available-modality conditions.

When the textual modality is absent, which corresponds to the scenarios of $\{a\}$, $\{v\}$, and $\{a, v\}$, our model delivers substantial improvements with accuracy increases of 7-12% over the best baseline results. This illustrates that our feature extraction and coding strategies can significantly enhance the acoustic and visual modalities compared to previous strategies. And when the textual modality is available ($\{t\}$, $\{a, t\}$, $\{v, t\}$), our model generally achieves superior performance. All of this demonstrates that the Bilateral Imagination Module is better able to retain the sentiment information, thereby enhancing the performance of sentiment prediction.

For the **MOSI** dataset, although our BiMIN slightly underperforms compared to TRF-Net and MPLMM under the specific $\{v, t\}$ condition, it still demonstrates superior overall performance. This is attributed to our BiMIN, which avoids attention-based or transformer mechanisms, focusing instead on both accuracy and efficiency. The average Acc and F1 score of our BiMIN exceeds 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.

Table 2
Quantitative results (%) for the case of the six available modality combinations. For example, "$\{t\}$" indicates that the textual modality is available while the acoustic and visual modalities are absent. "Avg" is the average performance of the six possible absent-modality combinations as shown in Table 1. "†" denotes results from (Guo et al., 2024). "*" denotes results for MCTN from (Pham et al., 2019). Other MCTN results are **not** reported in (Pham et al., 2019). **Bold**: best result. Underline: second best result. Higher values are better for all experimental results.

Datasets	Model	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{v,t\}$		$\{a,t\}$		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	79.3*	79.1*	53.1*	53.2*	76.8*	76.8*	76.4*	76.4*	64.12	60.86
	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**

On the **SIMS** dataset, our model slightly underperforms MPLMM and MPMM in terms of F1 scores under the $\{a\}$ and $\{v\}$ available-modality conditions, but our model surpasses them in terms of Acc. This suggests that although MPLMM and MPMM 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 dataset with Chinese contexts, and our model's strong performance on this dataset further demonstrates its effectiveness not only in English but also in Chinese contexts.

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, achieving notable enhancements in both Acc and F1 scores compared to the baselines. 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 optimal.

4.5. Supplementary Experiments

We propose BiMIN, a unified model designed to address the diverse combinations of available modalities in different scenarios. While its primary goal is to tackle the issue of absent modalities that the full-modal models cannot handle, BiMIN also performs effectively even when all three modalities are available. To further demonstrate the model's broad applicability across various scenarios, we conduct sentiment analysis experiments under full-modal conditions, using the original test set with all modalities available. As shown in Table 3, our model exhibits high and

Table 3
Quantitative results (%) for the full-modality availability case. "*" denotes results for MCTN from (Pham et al., 2019). Other MCTN results are **not** reported in (Pham et al., 2019). Parenthetically, the results on MPMM and MPLMM are **not** reported in (Guo et al., 2024). **Bold**: best result. Underline: second best result. Higher values are better for all experimental results.

Model	MOSI		SIMS		MOSEI	
	Acc	F1	Acc	F1	Acc	F1
MCTN	79.3*	79.1*	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**

stable performance across different datasets. This demonstrates that our BiMIN can flexibly adjust to the requirements of scenarios with varying available modalities.

4.6. Ablation Study on CMU-MOSEI

We perform ablation experiments to assess the roles of the Bilateral Auto-Encoder of the Bilateral Imagination Module and the Pre-trained Feature Extractors of the Modality Encoder Network within the BiMIN. Notably, in the experiments where the core and detail branches are absent (Bilateral Auto-Encoders are not utilized), Auto-Encoders are used as substitutes for Bilateral Auto-Encoders. For the experiments where the acoustic and visual modalities do not utilize Pre-trained Feature Extractors, we employ the same feature extraction approach as used in (Yu et al., 2020) and the same encoders as used in (Zhao et al., 2021).

Pre-trained Feature Extractors. As illustrated in Table 4, the experimental results clearly showcase 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 utilize 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 Feature Extractors is not limited to specific models but is also applicable across different model architectures. Regarding the extent of performance improvement, it is observed that in the presence of only the acoustic modality available (a), accuracy improves by 5% to 8%; in the presence of only the visual modality available ($\{v\}$), accuracy increases by 14% to 16%; and in the combined presence of only both acoustic and visual modalities available ($\{a, v\}$), accuracy gains rang from 11% to 15%.

Bilateral Auto-Encoder. As shown in Table 4, Experiments 1 and 2 utilize Auto-Encoders as the core architecture of the Modality Imagination Network, while Experiments 7 and 8 employ complete Bilateral Auto-Encoders as the backbone. Despite using the same Pre-trained Feature Extractors, it is observed that the average performance in Experiment 7 exceeds that of Experiment 1 by 3%, and Experiment 8 outperforms Experiment 2 by 3% as well. To further explore the necessity of the dual-branch structure in Bilateral Auto-Encoders, we compare 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 the experimental configurations, Experiment 8 stands out because it combines both acoustic and visual Pre-trained Feature Extractors and Bilateral Auto-Encoders. When we evaluate it across six different modality availability scenarios, Experiment 8 consistently delivers the best performance, with enhancements that are both balanced and stable.

Examining the second-best results, we see that models combined with acoustic and visual Pre-trained Feature Extractors show performance improvements when textual modality is absent. In contrast, models using Bilateral Auto-Encoders show performance improvements when textual modality is present. However, this does not mean that Bilateral Auto-Encoders can only improve performance with the presence of text modalities; they can also enhance performance in the absence of text modalities

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 acoustic and visual Feature Extractors **PE**, the Core Branch **CB** and the Detail Branch **DB** of the Bilateral Auto-Encoder. A "$\surd$" indicates that the module is 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	DB	$\{a\}$		$\{v\}$		$\{t\}$		$\{a,v\}$		$\{v,t\}$		$\{a,t\}$		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**

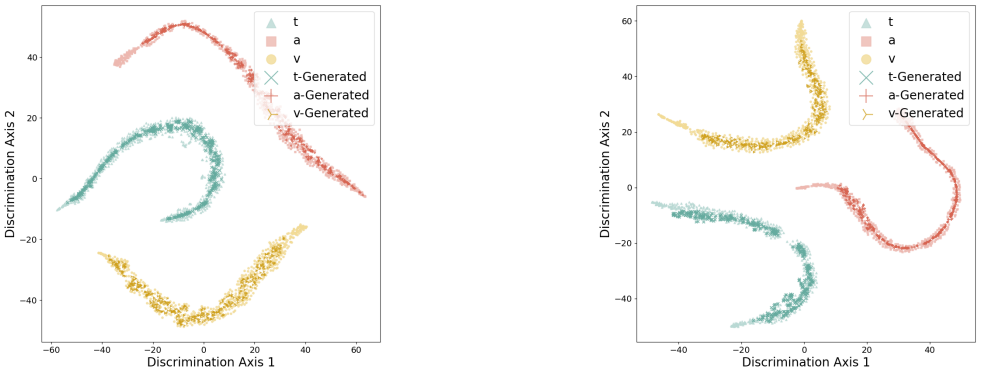
Table 5
On ablation experiments with Bilateral Auto-Encoders. We report the average performance for the six available modality combinations. Where "**MMIN**" and "**If-MMIN**" denote the original backbone network using the Auto-Encoder; "**MMIN(BiAE)**" and "**If-MMIN(BiAE)**" denotes that the Auto-Encoders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

Model	Acc↑	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

To further validate the performance advantages of the Bilateral Auto-Encoder, we conducted a controlled experiment under identical conditions. Specifically, we replace the Auto-Encoders in the MMIN and If-MMIN models with Bilateral Auto-Encoders and compared the results, as summarized in Table 5. The data reveal a clear trend: models equipped with Bilateral Auto-Encoders achieve significant improvements 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 show the Bilateral Auto-Encoders' ability to capture both coarse-grained core information and fine-grained detail information.

4.7. Capability Analysis

The core advantage of BiMIN lies in its unique Bilateral Imagination Module, which is specifically designed to generate representations for absent modalities. To empirically assess the effectiveness of this module, we carry out a series of carefully designed visualization experiments. As illustrated in Fig. 4, we utilize t-SNE (van der Maaten and Hinton, 2008) to compare the visualized distributions of the ground-truth multimodal embeddings (as shown by h in Fig. 1) with the multimodal embeddings generated by Bi-MMIN (denoted as $\hat{h}$ in Fig. 1). For the MOSEI test set, we randomly select 512 sets of multimodal data for analysis, while for the SIMS test set, we include 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 MOSEI dataset with English texts and the SIMS dataset with Chinese texts. Notably, the generation quality for all three modalities—textual, acoustic, and visual—is 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) **CMU-MOSEI** (b) **CH-SIMS**

Fig. 4: Visualisation of grand-truth multimodal embeddings and generated multimodal embeddings. For example, "**t**" denotes 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 an innovative sentiment analysis model designed to tackle modality uncertainty: the Bilateral Modality Imagination Network (BiMIN). This model comprises two key components: a Modality Encoder Network based on a pre-trained feature extractor, and a Bilateral Imagination Module based on the Bilateral Auto-Encoder. The Modality Encoder Network excels at extracting sentiment from the available modalities, thereby enhancing overall performance. Meanwhile, the Bilateral Imagination Module boosts the robustness of joint multimodal representations and improves prediction accuracy. Experimental results on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets demonstrate that BiMIN consistently outperforms baselines across various absent-modality combinations. Future research will concentrate on enhancing both the Modality Encoder Network and the Bilateral Imagination Module.

Acknowledgement

This study was supported by the Shanghai Service Industry Development Fund and the High Performance Computing Center of Shanghai University, and the Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600).

Data availability

In this work, we have used three publicly available datasets MOSI, MOSEI, SIMS and they can be downloaded from https://github.com/thuiar/MMSA.

Competing Interests Statement

The authors declared that they have no conflict of interest to this article.

Ethics approval

This article has never been submitted to more than one journal for simultaneous consideration. This article is original.

Consent to participate

The authors have approved this article before submission, including the names and order of authors.

References

Baltrušaitis, T., Robinson, P., Morency, L.P., 2016. Openface: an open source facial behavior analysis toolkit, in: 2016 IEEE winter conference on applications of computer vision, pp. 1–10.

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

Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S., 2014. COVAREP - A collaborative voice analysis repository for speech technologies, in: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 960–964.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

Du, C., Du, C., Wang, H., Li, J., Zheng, W.L., Lu, B.L., He, H., 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data, in: Proceedings of the 26th ACM international conference on Multimedia, pp. 108–116.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462.

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

Guo, Z., Jin, T., Zhao, Z., 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 1726–1736.

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

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations., pp. 1–15.

Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y., 2023. Multimodal prompting with missing modalities for visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14943–14952.

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

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

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

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

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605.

McFee, B., Raffel, C., Liang, D., Ellis, D.P.W., McVicar, M., Battenberg, E., Nieto, O., 2015. LibROSA: Audio and music signal analysis in python, in: Proceedings of the 14th Python in Science Conference, pp. 18–24.

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H., 2022. Expanding language-image pretrained models for general video recognition, in: European Conference on Computer Vision, pp. 1–18.

Pandey, G., Dukkipati, A., 2017. Variational methods for conditional multimodal deep learning, in: 2017 international joint conference on neural networks (IJCNN), pp. 308–315.

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

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, pp. 28492–28518.

Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.P., Hoque, E., 2020. Integrating multimodal information in large pretrained transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359–2369.

Tang, J., Li, K., Jin, X., Cichocki, A., Zhao, Q., Kong, W., 2021. 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, pp. 5301–5311.

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

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

Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558–6569.

Wang, L., Peng, J., Zheng, C., Zhao, T., Zhu, L., 2024. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Information Processing Management 61, 103675.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., Yang, K., 2020. 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, pp. 3718–3727.

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

Yuan, Z., Li, W., Xu, H., Yu, W., 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4400–4407.

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

Zadeh, A., Lim, Y.C., Morency, L.P., 2018a. Openface 2.0: Facial behavior analysis toolkit tadas baltrušaitis, in: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 59–66.

Zadeh, A., Zellers, R., Pincus, E., Morency, L.P., 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems 31, 82–88.

Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P., 2018b. 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, pp. 2236–2246.

Zeng, J., Liu, T., Zhou, J., 2022a. 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, pp. 1545–1554.

Zeng, J., Zhou, J., Liu, T., 2022b. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 2924–2934.

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters 23, 1499–1503.

Zhao, J., Li, R., Jin, Q., 2021. 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, pp. 2608–2618.

Zhao, T., Peng, J., Huang, Y., Wang, L., Zhang, H., Cai, Z., 2023. A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis. Appl. Intell. 53, 30455–30468.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.

Zuo, H., Liu, R., Zhao, J., Gao, G., Li, H., 2023. 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), pp. 1–5.