Pattern Recognition

Hierarchical Factorization-Guided Self-Distillation for Incomplete Multimodal Sentiment Analysis

--Manuscript Draft--

Manuscript Number:	PR-D-25-01766
Article Type:	VSI: WILD-VISION
Section/Category:	Various applications
Keywords:	Human sentiment understanding in videos; uncertain modality missing; representation factorization; multimodal representation; adversarial learning
Corresponding Author:	Dingkang Yang Fudan University Academy for Engineering and Technology CHINA
First Author:	Ziyang Liu
Order of Authors:	Ziyang Liu
	Mingcheng Li
	Yiquan Wang
	Dingkang Yang
Abstract:	Multimodal Sentiment Analysis (MSA) is an important research area that aims to understand and recognize human sentiment through multimodal data in videos. The complementary information provided by multimodal fusion promotes better sentiment analysis compared to utilizing only unimodal modalities. Nevertheless, in real-world applications, many unavoidable factors may lead to situations of uncertain modality missing, thus hindering the effectiveness of multimodal modeling and degrading the performance of the model. To this end, we propose a Hierarchical Factorization-Guided self-Distillation (HFGD) framework for the MSA task under uncertain missing modalities. Specifically, we propose a temporal saliency-based factorization mechanism that dynamically maximizes the mutual information between the representations based on temporal saliency, leading to the semantic-level factorization of sentimentrelevant and modality-exclusive representations. Moreover, we propose a distributional alignment-based factorization mechanism that further aligns the distributions between sentiment-relevant and modality-exclusive representations utilizing both contrastive and adversarial learning to produce robust joint multimodal representations. Eventually, a hierarchical self-distillation module is proposed to precisely reconstruct the missing features by multi-scale and progressive consistency supervision paradigms. Comprehensive experiments on three datasets indicate that HFGD has a significant performance advantage over previous methods with uncertain missing modalities and comparable performance with complete modalities.

Cover Letter

Dear editor:

We are submitting a manuscript entitled "**Hierarchical Factorization-Guided Self-Distillation for Incomplete Multimodal Sentiment Analysis**" to *Pattern Recognition* for publication. All authors have consulted the guide for authors, and we confirm that the manuscript has not been previously published, in whole or in part, and is not under consideration by any other journal.

This manuscript focuses on the challenge of uncertain modality missing in multimodal sentiment analysis, which has received widespread attention in the video processing area. Although existing methods improve the MSA performance with missing modalities, they have the following limitations: (i) Implementing complex feature interactions for incomplete modalities leads to a large amount of information redundancy and cumulative errors. (ii) Static and coarse-grained representation reconstruction paradigms lead to imprecise missing feature reconstruction. This leads to their inability to precisely reconstruct the missing sentiment semantics. To this end, we propose a Hierarchical Factorization-Guided self-Distillation (HFGD) framework to address the aforementioned issues. HFGD consists of three core components: Temporal Saliency-based Factorization (TSF) mechanism, Distributional Alignment-based Factorization (DAF) mechanism, and Hierarchical Self-Distillation (HSD) module. TSF dynamically maximizes the mutual information among the representations based on temporal saliency, which facilitates the semantic-level factorization of collaborative representation and modality-exclusive representations. DAF enhances the alignment of distributions between collaborative representation and modality-exclusive representations by jointly using contrastive learning and adversarial learning to generate robust joint multimodal representations. HSD precisely reconstructs the missing sentiment semantics by multi-scale and progressive consistency supervision. Based on these components, HFGD has a significant performance advantage over previous methods with uncertain missing modalities and comparable performance with complete modalities on three multimodal benchmarks

More importantly, our framework can be applied to diverse real-world application scenarios, accommodating uncertain modalities missing due to many inevitable factors, such as privacy, device, or security constraints. The topic addressed is suitable for this journal, and the findings would be helpful for the development of the video technology community. We deeply appreciate your consideration of our manuscript, and we look forward to receiving comments from the reviewers.

Best regards

Dingkang Yang
dkyang20@fudan.edu.cn
Academy for Engineering & Technology, Fudan University

Title:
Hierarchical Factorization-Guided Self-Distillation for Incomplete Multimodal Sentiment Analysis

Author names and affiliations:
Ziyang Liu (ziyannn@yeah.net)[a], Mingcheng Li (mingchengli21@m.fudan.edu.cn)[b], Yiquan Wang[c], Dingkang Yang (lihuazhang@fudan.edu.cn)[b]

[a]School of future science and engineering, Soochow University, Suzhou ,215000,China
[b]Academy for Engineering and Technology, Fudan University, Shanghai, 200433, China
[c]College of Mathematics and System Science, Xinjiang University, Urumqi, 830046, China

Corresponding author:
Dingkang Yang. Email address: dkyang20@fudan.edu.cn

Highlights

- A Hierarchical Factorization-Guided self-Distillation framework is presented to alleviate the severe modality missing problem in multimodal sentiment analysis.

- The proposed temporal saliency-based factorization mechanism that dynamically maximizes the mutual information among the representations based on temporal saliency. This approach facilitates the semantic-level factorization of collaborative representations and modality-exclusive representations.

- We propose a distributional alignment-based factorization mechanism that enhances the alignment of distributions between sentiment-relevant and modality-exclusive representations. This alignment is accomplished by jointly using contrastive learning and adversarial learning to generate robust joint multimodal representations.

- A hierarchical self-distillation module is proposed to precisely reconstruct the missing sentiment semantics by multi-scale and progressive consistency supervision.

- Comprehensive experiments on three datasets indicate that our method has a significant performance advantage over previous methods with uncertain missing modalities and comparable performance with complete modalities.

Hierarchical Factorization-Guided Self-Distillation for Incomplete Multimodal Sentiment Analysis

Ziyang Liu[a], Mingcheng Li[b], Yiquan Wang[c], Dingkang Yang[b,*]

[a]*School of future science and engineering, Soochow University, Suzhou ,215000,China*
[b]*Academy for Engineering and Technology, Fudan University, Shanghai, 200433, China*
[c]*College of Mathematics and System Science, Xinjiang University, Urumqi, 830046, China*

Abstract

Multimodal Sentiment Analysis (MSA) is an important research area that aims to understand and recognize human sentiment through multimodal data in videos. The complementary information provided by multimodal fusion promotes better sentiment analysis compared to utilizing only unimodal modalities. Nevertheless, in real-world applications, many unavoidable factors may lead to situations of uncertain modality missing, thus hindering the effectiveness of multimodal modeling and degrading the performance of the model. To this end, we propose a Hierarchical Factorization-Guided self-Distillation (HFGD) framework for the MSA task under uncertain missing modalities. Specifically, we propose a temporal saliency-based factorization mechanism that dynamically maximizes the mutual information between the representations based on temporal saliency, leading to the semantic-level factorization of collaborative representations and modality-exclusive representations. Moreover, we propose a distributional alignment-based factorization mechanism that further aligns the distributions between collaborative representation and modality-exclusive representations utilizing both contrastive and adversarial learning to produce robust joint multimodal representations. Eventually, a hierarchical self-distillation module is proposed to precisely reconstruct the missing features by multi-scale

*Corresponding author.
Email address: ziyannn@yeah.net, mingchengli21@m.fudan.edu.cn, dkyang20@fudan.edu.cn (Dingkang Yang)

and progressive consistency supervision paradigms. Comprehensive experiments on three datasets indicate that HFGD has a significant performance advantage over previous methods with uncertain missing modalities and comparable performance with complete modalities.

Keywords: Human sentiment understanding in videos, uncertain modality missing, representation factorization, multimodal representation, adversarial learning

---

Multimodal sentiment analysis (MSA) has garnered significant attention in recent years. Unlike traditional unimodal-based emotion recognition tasks [1, 2], MSA leverages multiple modalities, including language, audio, and visual, to understand and recognize human emotions [3, 4, 5, 6]. Previous research has indicated that the fusion of complementary semantics from different modalities facilitates the generation of refined joint multimodal representations [5**?**]. To date, MSA research has primarily operated under the presumption that all modalities are available during both the training and inference stages [7, 8, 9, 10, 11, 12, 13, 14]. Nevertheless, in real-world applications, modalities may be missing due to security concerns, background noises, sensor limitations, and so on. Ultimately, these incomplete multimodal data significantly hinder the performance of MSA. For example, uncertain modality missing in real-world applications leads to incorrect sentiment recognition, as illustrated in Figure 1. Consequently, effectively addressing the missing modality problem is essential for enhancing the robustness and stability of MSA systems.

1. Introduction

In recent years, many studies [15, 16, 17, 18, 19, 20, 21] attempt to address the problem of missing modalities in MSA. For example, SMIL [22] estimates the latent features of the missing modality data via Bayesian Meta-Learning. GCNet [16] utilizes both speaker and temporal information in conversations to derive discriminative representations from data with incomplete modalities.

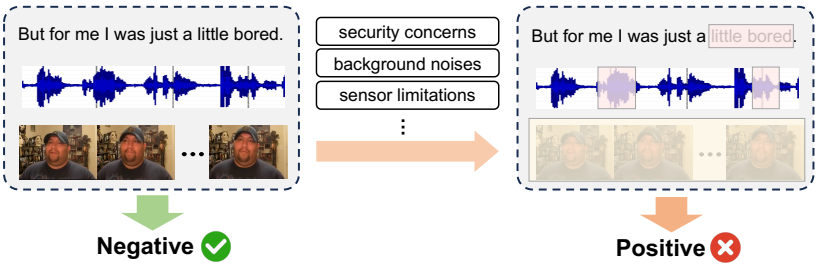


Figure 1: Examples of uncertain missing modalities. Traditional model outputs correct prediction when inputting the sample with complete modalities, but incorrectly predict when modalities are missing. The modality- missing cases are categorized into two types: (1) intra-modality missingness (*i.e.*, pink areas) and (2) inter-modality missingness (*i.e.*, yellow area).

Additionally, incomplete multimodality learning has achieved some success in other tasks [23, 24]. For example, MECOM [23] proposes a meta-completion approach that utilizes cross-modality attention and decoupled reconstruction to explicitly complete missing modality. 3D-IMMC [24] presents an incomplete multi-modal 3D shape clustering method with cross mapping and dual adaptive fusion to mitigate the negative impact of missing modality instances on multi-modal 3D shapes. However, these methods are constrained by the following factors: (i) Implementing complex feature interactions for incomplete modalities leads to a large amount of information redundancy and cumulative errors, resulting in ineffective extraction of sentiment semantics and their distributions. (ii) Static and coarse-grained representation reconstruction paradigms lead to imprecise missing semantic reconstruction and nonrobust joint multimodal representations.

To address the above issues, we propose a Hierarchical Factorization-Guided self-Distillation (HFGD) framework for the MSA task under uncertain missing modalities. HFGD has three core contributions: (i) we propose a temporal saliency-based factorization mechanism that dynamically maximizes the mutual information among the representations based on temporal saliency. This approach facilitates the semantic-level factorization of collaborative representations and modality-exclusive representations. (ii) Additionally, we propose

a distributional alignment-based factorization mechanism that enhances the alignment of distributions between collaborative representation and modality-exclusive representations. This alignment is accomplished by jointly using contrastive learning and adversarial learning to generate robust joint multimodal representations. (iii) Ultimately, a hierarchical self-distillation module is proposed to precisely reconstruct the missing sentiment semantics by multi-scale and progressive consistency supervision. Based on these components, HFGD has a significant performance advantage over previous methods with uncertain missing modalities and comparable performance with complete modalities on three multimodal benchmarks.

2. Related Work

2.1. Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) seeks to comprehend and analyze human sentiment by utilizing diverse modalities. Unlike conventional single-modality sentiment recognition, MSA poses greater challenges owing to the intricate nature of processing and analyzing heterogeneous data across modalities. Mainstream studies in MSA [25, 26, 27, 11, 28, 29, 30, 31] focus on designing complex fusion paradigms and interaction mechanisms to improve MSA performance. For instance, CubeMLP [27] employed three distinct multi-layer perceptron units for feature amalgamation along three axes. However, these methods assume the availability of complete modalities, making them impractical for real-world deployment. The primary approaches addressing missing modalities in MSA can be categorized into two paradigms: (i) generative methods [15, 16, 17] and (ii) joint learning methods [18, 19, 20]. Generative methods aim to regenerate missing features and semantics within modalities by leveraging the distributions of available modalities. For instance, MVAE [15] addressed modality missingness through a semi-supervised multi-view deep generative framework. On the other hand, joint learning methods focus on deriving cohesive joint multimodal representations based on inter-modality correlations.

For instance, CorrKD [32] proposes a correlation-decoupled knowledge distillation framework that utilizes cross-sample, cross-category, and cross-response correlations to generate robust joint multimodal representations for mitigating the modality missing problem in MSA. TATE [20] incorporated a tag encoding module to direct the network's attention toward missing modalities. However, these methods lack effective capture and reconstruction of the sentiment semantics and distribution within the modality, resulting in non-robust joint multimodal representations and poor performance. In contrast, our fine-grained factorization schema sufficiently exploits and aligns the sentiment semantics in the modality, thus precisely recovering the missing features.

2.2. Factorized Representation Learning

The core criterion for representation factorization learning is to separate representations with distinct semantics and distributions in the data, thus facilitating the model to better capture the inherent information in the data and produce more valuable modality representations. Previous works on factorized representation learning are mainly based on auto-encoders and generative adversarial networks. FactorVA [33] was proposed to implement factorization based on the property that representations are factorial and independent in dimension. Wu *et al.* [34] introduced a disentangled variational approach that aligns correlations across variations in different modalities for heterogeneous face matching. In recent years, factorization learning has been progressively applied to MSA tasks. For example, FDMER [8] utilized consistency and discreteness constraints between modalities to disentangle modalities into modality-invariant and modality-private features. DMD [11] disentangled each modality into modality-independent and modality-exclusive representations and then implemented a knowledge distillation strategy among the representations with dynamic graphs. MEA [12] views modality factorization as the capture and mining of modal-agnostic and modal-exclusive information. Based on this foundation, this work proposes the homogeneous and heterogeneous graph fusion mechanism to multifacetedly realize knowledge transfer and information flow between

modalities, improving the dilemma of the modality heterogeneity. Although these works have made some progress in MSA, there are still some limitations: (i) only factorizing distinct representations at the modality level without considering sentimentally beneficial and relevant representations. (ii) The supervision of the factorization process is coarse-grained and inadequate. In contrast, the proposed method accurately factorizes collaborative representations at the semantic level through temporal saliency modeling based on mutual information maximization. Furthermore, contrastive learning and adversarial learning are used to align the sentiment semantics in the representation at the distributional level in order to produce robust joint multimodal representations.

2.3. Knowledge Distillation

Knowledge distillation leverages supervisory signals from a pre-trained teacher network to facilitate the training of a student network. There are generally two categories of knowledge distillation methods: distillation from intermediate features [35] and distillation from logits [36, 37]. Many studies [38, 39, 40] have employed knowledge distillation for MSA tasks with missing modalities. These approaches focus on transferring knowledge to the student network using a teacher network that contains richer and more comprehensive semantics, thus addressing the missing modality problem to some extent. For instance, LCKD [40] proposed a cross-modal knowledge distillation method to dynamically extract beneficial knowledge from important modalities to mitigate the negative effects caused by missing modalities. KD-Net [39] employed the teacher network with complete modalities to supervise the student network with missing modalities at both the feature and logits levels. Despite the progress made by the aforementioned methods, there are still some limitations (i) The teacher network brings a large memory footprint. (ii) Only a unidirectional knowledge transfer from the teacher network to the student network makes it difficult to sufficiently reconstruct missing features and semantics. (iii) Supervision between networks is inadequate and lacks consideration of semantic and distributional alignment. To solve these issues, we design a hierarchical self-distillation mod-

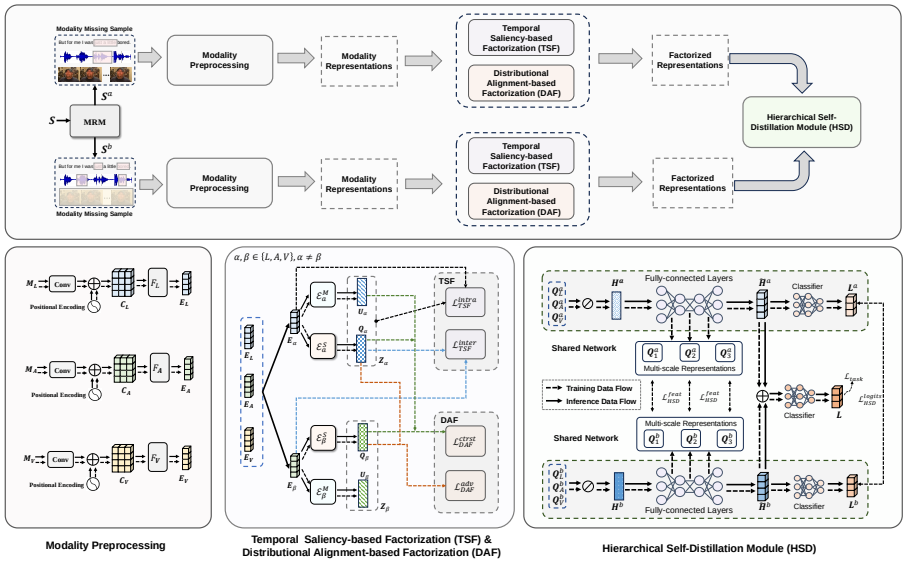


Figure 2: The structure of our HFGD, which consists of three core components: Temporal Saliency-based Factorization (TSF) mechanism, Distributional Alignment-based Factorization (DAF) mechanism, and Hierarchical Self-Distillation (HSD) module.

ule that drives a single network to implement bi-directional knowledge transfer with low overhead to precisely reconstruct missing information and generate valuable joint multimodal representations.

3. Methodology

3.1. Problem Formulation

Each sample is a multimodal video segment containing three distinct modalities, denoted as $\boldsymbol{S} = [\boldsymbol{X}_L, \boldsymbol{X}_A, \boldsymbol{X}_V]$, where $\boldsymbol{X}_L \in \mathbb{R}^{T_L \times d_L}, \boldsymbol{X}_A \in \mathbb{R}^{T_A \times d_A}$, and $\boldsymbol{X}_V \in \mathbb{R}^{T_V \times d_V}$ denote language, audio, and visual modalities, respectively. $\xi = \{L, A, V\}$ denotes the set of modality types. $T_m(\cdot)$ and $d_m(\cdot)$ represent the sequence length and the embedding dimension, where $m \in \xi$. To realistically simulate the uncertain modality missing phenomenon in real-world applications, we set up two modality missing cases: (1) *intra-modality missingness*, which indicates some frame-level features in the modality sequences are missing. (2)

inter-modality missingness, which indicates some modalities are entirely missing. Our objective is to recognize sentiments at the utterance level by leveraging multimodal data that includes missing modalities.

3.2. Overall Framework

Figure 2 shows the architecture of the proposed HFGD and its main workflows in the training and testing phases are described in detail as follows.

Training Phase. The workflow of the training phase consists of several steps: (i) For each sample $\boldsymbol{S}$, we generate two heterogeneous modality missing samples $\boldsymbol{S}^a$ and $\boldsymbol{S}^b$ with the Modality Random Missing (MRM) strategy. MRM simultaneously performs intra-modality missingness and inter-modality missingness. $\boldsymbol{S}^a$ and $\boldsymbol{S}^b$ are fed into two shared networks in parallel and symmetrically. (ii) We input each sample into the proposed temporal saliency-based factorization and distributional alignment-based factorization mechanisms to factorize each modality into a collaborative representation $\boldsymbol{Q}_m^k$ and a modality-exclusive representation $\boldsymbol{U}_m^k$, where $m \in \xi$ and $k \in \{a, b\}$. (iii) We concatenate three collaborative representations of both samples to obtain joint multimodal representations $\boldsymbol{H}^a$ and $\boldsymbol{H}^b$, which are subsequently fed into the hierarchical self-distillation module to learning feature-level and logits-level consistency. (iv) $\tilde{\boldsymbol{H}}^a$ and $\tilde{\boldsymbol{H}}^b$ are added and fed into the fully-connected layers for sentiment recognition.

Inference Phase. In the inference phase, the workflow consists of the following steps: (i) Given an input sample with missing modality, we make a copy of it to obtain two identical samples $\boldsymbol{S}^a$ and $\boldsymbol{S}^b$. (ii) We input these two identical samples into the framework in parallel and symmetrically, and after passing through encoders, we obtain $\boldsymbol{Q}_m^a$, $\boldsymbol{U}_m^a$, $\boldsymbol{Q}_m^b$ and $\boldsymbol{U}_m^b$ with $m \in \{L, A, V\}$. (iii) The collaborative representations are fed into the HSD to obtain $\tilde{\boldsymbol{H}}^a$ and $\tilde{\boldsymbol{H}}^b$, which are summed up and fed into the fully-connected layer for sentiment recognition.

3.3. Temporal Saliency-based Factorization

In modeling and fusion of multimodal sequences for MSA tasks, distribution gaps and information redundancy caused by modality heterogeneity cause

models to produce sentiment-irrelevant and ambiguous semantics. Additionally, the cases of missing uncertain modalities present a significant challenge to effectively extracting valuable affective information. Although previous studies have already made some progress in MSA with missing modalities [18, 19], they focused on implementing complex feature interactions for incomplete modalities. This paradigm leads to a large amount of information redundancy and cumulative error, which results in low-quality sentiment semantic refinement. Therefore, we propose a Temporal Saliency-based Factorization (TSF) mechanism to capture sentiment semantics in modality representations. The central idea is to decompose each modality into two different types of representations: (1) *Collaborative representation*, which captures the overall sentiment semantics of the sample. This representation is independent of any particular modality, is consistent across all modalities for the same subject, and is resilient to situations where modalities may be missing. (2) *Modality-exclusive representation*, which reflects information unique to each modality that is unrelated to the task, capturing the specific characteristics inherent to that modality.

As shown in Figure 2, our framework receives two heterogeneous modality missing samples $\boldsymbol{S}^a$ and $\boldsymbol{S}^b$ as input in parallel and symmetrically. For brevity, here we denote a particular sample of the input to the network with $\boldsymbol{S} = [\boldsymbol{X}_L, \boldsymbol{X}_A, \boldsymbol{X}_V]$ with modality number $M = 3$. We define the modalities in sample $\boldsymbol{S}$ as $\boldsymbol{X}_\alpha$ with $\alpha \in \xi$. Each modality is processed through a 1D temporal convolutional layer with kernel size 3×3. Then, positional embeddings [41] are added to generate the preliminary representations, which are expressed as $\boldsymbol{C}_\alpha = \boldsymbol{W}_{3\times3}(\boldsymbol{X}_\alpha) + PE(T_\alpha, d) \in \mathbb{R}^{T\times d}$. The $\boldsymbol{C}_\alpha$ is input into a Transformer [41] encoder $\mathcal{F}_\alpha(\cdot)$ to obtain $\boldsymbol{E}_\alpha = \mathcal{F}_\alpha(\boldsymbol{C}_\alpha) \in \mathbb{R}^{T\times d}$. The $\boldsymbol{E}_\alpha$ is the low-level modality representation of the modality α. Each modality representation is factorized into two representations: a Collaborative representation $\boldsymbol{Q}_\alpha$ using a sentiment encoder $\mathcal{E}^S_\alpha$, represented as $\boldsymbol{Q}_\alpha = \mathcal{E}^S_\alpha(\boldsymbol{E}_{\boldsymbol{\alpha}})$, and a modality-exclusive representation $\boldsymbol{U}_\alpha$ using a modality encoder $\mathcal{E}^M_\alpha$, represented as $\boldsymbol{U}_\alpha = \mathcal{E}^M_\alpha(\boldsymbol{E}_{\boldsymbol{\alpha}})$. $\boldsymbol{Z}_\alpha$ is the concatenation of $\boldsymbol{Q}_\alpha$ and $\boldsymbol{U}_\alpha$. The encoders $\mathcal{E}^S_\alpha(\cdot)$ and $\mathcal{E}^M_\alpha(\cdot)$ are implemented using multi-layer perceptrons with ReLU activation functions. To

achieve adequate factorization, we are required to ensure that collaborative representations have sentiment-discriminative power and that modality-exclusive representations capture the remaining sentiment-irrelevant information within the modality. Consequently, we propose a Temporal Saliency-based Factorization (TSF) mechanism to achieve sufficient factorization of modalities. The core ideas of TSF are (1) the collaborative representation of a modality is expected to have similar semantics to the low-level representations of other modalities, and (2) the collaborative representation and modality-exclusive representations factorized by a modality are semantically similar to the low-level representations of this modality. Moreover, temporal saliency-based Mutual Information (MI) is used to measure the semantic consistency among representations. Specifically, given any two modalities $\boldsymbol{X}_\alpha$ and $\boldsymbol{X}_\beta$, their representations $\boldsymbol{E}_\alpha$ and $\boldsymbol{E}_\beta$ are uniformly sliced into $N = 10$ non-overlapping segments in the temporal dimension. Each segment is denoted as $\boldsymbol{E}_\alpha^i$ and $\boldsymbol{E}_\beta^i$ with $i \in \{1, 2, \cdots, 10\}$, and is assigned a learnable temporal saliency weight ω^i to measure its contribution to the sentiment recognition. Immediately, a series of 1D convolutional layers are introduced to integrate the temporal information in $\boldsymbol{E}_{\alpha/\beta}^i$ and project the sequence lengths into dimensions matching $\boldsymbol{Q}_\alpha$ and $\boldsymbol{Z}_\alpha$, respectively. We define $MI(\boldsymbol{x}, \boldsymbol{y})$ as the MI between $\boldsymbol{x}$ and $\boldsymbol{y}$. The TSF loss is expressed as a weighted average of MI, denoted as:

$$\mathcal{L}_{TSF}^{inter} = -\frac{1}{M(M-1)} \frac{1}{N} \sum_{\alpha \in \xi} \sum_{\beta \in \xi, \beta \neq \alpha} \sum_{i}^{N} \omega_\beta^i MI(\boldsymbol{Q}_\alpha, \boldsymbol{E}_\beta^i), \tag{1}$$

$$\mathcal{L}_{TSF}^{intra} = -\frac{1}{M} \frac{1}{N} \sum_{\alpha \in \xi} \sum_{i}^{N} \omega_\alpha^i MI(\boldsymbol{Z}_\alpha, \boldsymbol{E}_\alpha^i). \tag{2}$$

To estimate the mutual information between representations, we define $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ be two random variables, and $P(\boldsymbol{x})$ and $P(\boldsymbol{y})$ are their marginal probability distributions. $P(\boldsymbol{x}, \boldsymbol{y})$ is the joint probability distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$. The MI can be defined as the relative entropy of the product of the joint probability distribution of two variables and their respective marginal probabil-

ity distributions, expressed as

$$MI(\boldsymbol{x}, \boldsymbol{y}) = \int_{\boldsymbol{\mathcal{X}}} \int_{\boldsymbol{\mathcal{Y}}} P(\boldsymbol{x}, \boldsymbol{y}) \log \left(\frac{P(\boldsymbol{x}, \boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})} \right) d\boldsymbol{x} d\boldsymbol{y}. \tag{3}$$

We only need to maximize the mutual information without focusing on its precise value. We adopt the mutual information estimation based on the Jensen-Shannon divergence, which has proven to be stable [42], denoted as follows:

$$\begin{aligned} MI(\boldsymbol{x}, \boldsymbol{y}) &= \hat{I}_{\theta}^{(\mathrm{JSD})}(\boldsymbol{x}, \boldsymbol{y}) \\ &= \mathbb{E}_{P(\boldsymbol{x},\boldsymbol{y})}[-sp(-\mathcal{F}_{\theta}(x, y)] \\ &- \mathbb{E}_{P(\boldsymbol{x})P(\boldsymbol{y})}[sp(\mathcal{F}_{\theta}(x, y)], \end{aligned} \tag{4}$$

where $sp(w) = \log(1 + e^w)$ and $\mathcal{F}_{\theta} : \boldsymbol{x} \times \boldsymbol{y} \to \mathbb{R}$ is a deep neural network of parameters θ called the statistics network. Consequently, the loss of temporal saliency-based factorization is formatted as follows:

$$\mathcal{L}_{TSF} = \mathcal{L}_{TSF}^{inter} + \mathcal{L}_{TSF}^{intra}. \tag{5}$$

3.4. Distributional Alignment-based Factorization

Although the TSF factorizes modality to some extent into two heterogeneous representations that contain distinct semantics, their distributions are unaligned, leading to an insufficient feature separation. To this end, we propose a Distributional Alignment-based Factorization (DAF) mechanism that further aligns the distributions between the two factorized representations by jointly employing contrastive and adversarial learning.

Given two arbitrary modalities, we narrow the distance between the collaborative representations of both modalities and enlarge the distance between the collaborative representation and the modality-exclusive representation of the same modality. The contrastive learning loss is denoted as:

$$\mathcal{L}_{DAF}^{ctrst} = \frac{1}{M(M-1)} \sum_{\alpha \in \xi} \sum_{\beta \in \xi, \beta \neq \alpha} \mathcal{D}(\boldsymbol{Q}_{\alpha}, \boldsymbol{Q}_{\beta})^2 + \max\{0, \eta - \mathcal{D}(\boldsymbol{Q}_{\alpha}, \boldsymbol{U}_{\alpha})\}^2 \tag{6}$$

where $\mathcal{D}(a, b) = \|a - b\|_2$, $\|\cdot\|_2$ represents ℓ_2 norm function, and η is the predefined distance boundary. Furthermore, adversarial learning is employed to

align the distributions between any two collaborative representations produced by factorization. Our framework tries to generate aligned collaborative representations to mislead the discriminator $\mathbb{D}$, which distinguishes between the two representations. In practice, $\mathbb{D}$ consists of a fully connected layer network. The adversarial learning loss is formatted as:

$$\mathcal{L}_{DAF}^{adv} = \frac{1}{M(M-1)} \sum_{\alpha \in \xi} \sum_{\beta \in \xi, \beta \neq \alpha} log(1 - \mathbb{D}(\boldsymbol{Q}_\alpha)) + log(\mathbb{D}(\boldsymbol{Q}_\beta)). \tag{7}$$

Consequently, the loss of distributional alignment-based factorization is expressed as follows:

$$\mathcal{L}_{DAF} = \mathcal{L}_{DAF}^{ctrst} + \mathcal{L}_{DAF}^{adv}. \tag{8}$$

3.5. Hierarchical Self-Distillation Module

Conventional knowledge distillation techniques for handling missing modalities utilize complete-modality teacher networks to direct the training of missing-modality student networks. These methods are hindered by several drawbacks, such as the necessity for high-performing teacher networks, substantial training expenditures, and the static and coarse-grained information transfer [38, 39]. To address the above issues, we propose a Hierarchical Self-Distillation (HSD) module to progressively learn the representation consistency and recover the missing sentiment semantics through a hierarchical self-distillation paradigm. Specifically, HSD utilizes bidirectional knowledge transfer within a single network to achieve feature and logits consistency constraints between two heterogeneous modality missing samples. This learning paradigm alleviates the unidirectional dependence on knowledge [43] and offers two main benefits: transferring knowledge from more to fewer modalities aids in recovering lost information of the missing modalities while transferring knowledge in the opposite direction enhances beneficial information. In summary, HSD can facilitate the model in producing more robust joint multimodal representations.

As shown in Figure 2, the representations $\boldsymbol{Q}_m^a \in \mathbb{R}^{T \times d}$ and $\boldsymbol{Q}_m^b \in \mathbb{R}^{T \times d}$ of three modalities are concatenated to obtain joint multimodal representations $\boldsymbol{H}^a \in \mathbb{R}^{T \times 3d}$ and $\boldsymbol{H}^b \in \mathbb{R}^{T \times 3d}$. Two representations $\boldsymbol{H}^a$ and $\boldsymbol{H}^b$ are fed into shared networks in parallel and symmetrically to implement the self-distillation

paradigm. Specifically, the fully-connected layers are utilized to refine the joint multimodal representation $\boldsymbol{H}^k \in \mathbb{R}^{T\times 3d}$ with $k \in \{a, b\}$, yielding $\tilde{\boldsymbol{H}}^k \in \mathbb{R}^{T\times 3d}$. Moreover, we obtain the intermediate multi-scale representations of all layers, denoted as $\boldsymbol{H}_1^k \in \mathbb{R}^{T\times 2d}$, $\boldsymbol{H}_2^k \in \mathbb{R}^{T\times d}$, and $\boldsymbol{H}_3^k \in \mathbb{R}^{T\times 2d}$. For the above five features, we implement the element-wise sum operation of the same scale to obtain $\boldsymbol{C}_1^k \in \mathbb{R}^{T\times 3d}$, $\boldsymbol{C}_2^k \in \mathbb{R}^{T\times 2d}$, and $\boldsymbol{C}_3^k \in \mathbb{R}^{T\times d}$, which are utilized in the subsequent computation and knowledge transfer. We supervise the consistency of the joint multimodal representation of two heterogeneous modality missing samples by feature-level and logits-level distillation.

Feature-level Distillation. To accurately measure the holistic discrepancy between the distributions of the two representations, we utilize the Wasserstein distance [44]. It is defined as the Optimal Mass Transportation (OMT) problem [45], which can nicely handle complex distributions in multiple fields [46]. The key principle of Wasserstein distance is to measure the distances between the two distributions by minimizing the smallest average transportation cost required to convert one distribution to the other. Define two probability distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ on the probability spaces $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$. The Wasserstein distance between $\boldsymbol{P}$ and $\boldsymbol{Q}$ is defined as:

$$W_p(\boldsymbol{P}, \boldsymbol{Q}) = \left(\inf_{\gamma \in \Gamma(\boldsymbol{P}, \boldsymbol{Q})} \int_{\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Y}}} \|\boldsymbol{x} - \boldsymbol{y}\|^p d\gamma(\boldsymbol{x}, \boldsymbol{y}) \right)^{1/p}, \tag{9}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\Gamma(\boldsymbol{P}, \boldsymbol{Q})$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are P and Q, and $p \geq 1$ is the order of the Wasserstein distance. Specifically, we adopt Wasserstein distance as a non-parametric metric to measure the discrepancy between each pair of same-scale feature representations, *i.e.*, $\boldsymbol{C}_j^a$ and $\boldsymbol{C}_j^b$ with scale $j \in \{1, 2, 3\}$. The feature distillation loss is represented as:

$$\mathcal{L}_{HSD}^{feat} = \sum_{j=1}^{3} W_p(\boldsymbol{C}_j^a, \boldsymbol{C}_j^b). \tag{10}$$

Logits-level Distillation. To reduce the distributional discrepancy between the logits, we create soft labels to guide the learning process. In the training phase, representations $\tilde{\boldsymbol{H}}^a$ and $\tilde{\boldsymbol{H}}^b$ pass through fully connected layers and

softmax layer to get logits $\boldsymbol{L}^a$ and $\boldsymbol{L}^b$. We utilize the Jensen-Shannon (JS) divergence as the measure of discrepancy. The JS divergence is a symmetric measure of similarity between two probability distributions. It is based on the Kullback-Leibler (KL) divergence and quantifies the divergence by averaging the KL divergence between each distribution and their mean. JS divergence is bounded between 0 and 1, making it a widely used metric for evaluating distributions. The loss associated with logits distillation is defined as follows:

$$\mathcal{L}_{HSD}^{logits} = \mathcal{D}_{JS}(\boldsymbol{L}^a||\boldsymbol{L}^b)) = \frac{1}{2}(\mathcal{D}_{KL}(\boldsymbol{L}^a||\boldsymbol{M})) + \mathcal{D}_{KL}(\boldsymbol{L}^b||\boldsymbol{M})), \tag{11}$$

where $\boldsymbol{L}^a$ and $\boldsymbol{L}^a$ are logits obtained from features $\tilde{\boldsymbol{H}}^a$ and $\tilde{\boldsymbol{H}}^b$ through the fully connected layer, and $\boldsymbol{M}$ is the average distribution of $\boldsymbol{L}^a$ and $\boldsymbol{L}^b$. Ultimately, the loss of HSD is denoted as:

$$\mathcal{L}_{HSD} = \mathcal{L}_{HSD}^{feat} + \mathcal{L}_{HSD}^{logits}. \tag{12}$$

3.6. Optimization Objectives

The overall optimization objective $\mathcal{L}_{total}$ is expressed as $\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{TSF} + \mathcal{L}_{DAF} + \mathcal{L}_{HSD}$, where $\mathcal{L}_{task}$ is the standard cross-entropy loss.

4. Experiments

4.1. Datasets and Evaluation Metrics

Our experiments are conducted on three MSA benchmarks: MOSI [47], MOSEI [48], and IEMOCAP [49]. These experiments are carried out in a word-aligned setting. The MOSI dataset, which is used for MSA, contains 2,199 video segments, which are split into 1,284 training clips, 229 validation clips, and 686 testing clips. The MOSEI dataset includes a total of 22,856 video segments, divided into 16,326 samples for training, 1,871 samples for validation, and 4,659 samples for testing. In the MOSI and MOSEI, each sample was labeled with sentiment scores ranging from -3 (strongly negative) to +3 (strongly positive) For the MOSI and MOSEI datasets, we employ two evaluation metrics: Mean Absolute Error (MAE) and F1 score for positive/negative classification.

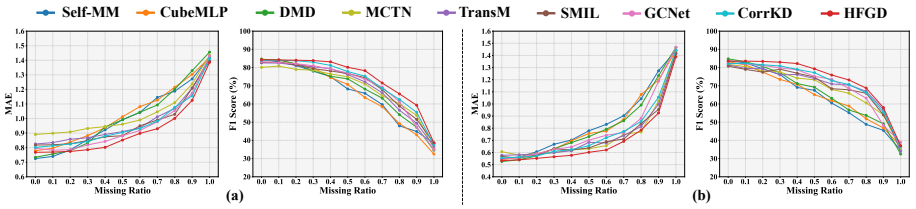


Figure 3: Comparison results of various missing ratios on (a) MOSI and (b) MOSEI.

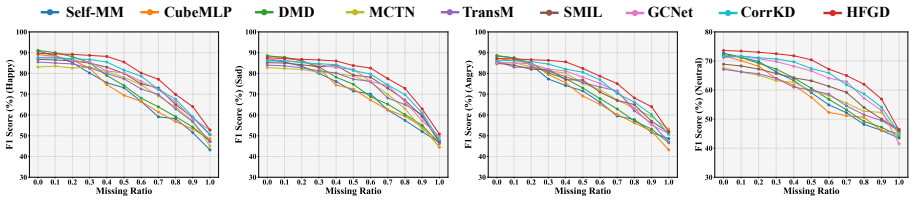


Figure 4: Comparison results of various missing ratios on IEMOCAP. We report on the F1 score evaluation metric for the happy, sad, angry, and neutral categories.

The IEMOCAP dataset consists of 4,453 video clip samples, which are divided into 2,717 samples for training, 798 samples for validation, and 938 samples for testing. For a fair comparison with previous works [25, 50], we focus on happy, sad, angry, and neutral emotions for MSA tasks. The evaluation metric is the F1 score.

4.2. Implementation Details

Here, we describe the feature extraction process for different modalities and the detailed experimental setup.

Feature Extraction. The Glove embedding [51] is used to convert the video transcripts to obtain a 300-dimensional vector for the language modality. For the audio modality, we employ the COVAREP toolkit [52] to extract 74-dimensional acoustic features, including 12 Mel-frequency cepstral coefficients (MFCCs), voiced/unvoiced segmenting features, and glottal source parameters. For the visual modality, we utilize the Facet [53] to capture 35 facial action units, which track facial movements associated with emotional expressions. The usage of the above multimodal features aligns most methods for fair and intuitive compar-

Table 1: Comparison results under inter-modality missingness cases on MOSI and MOSEI. The MAE metric is reported.

Datasets	Models	Testing Conditions							
		$\{l\}$	$\{a\}$	$\{v\}$	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	Avg.	$\{l,a,v\}$
MOSI	Self-MM [7]§	1.008	1.396	1.452	0.993	0.961	1.331	1.190	**0.725**
	CubeMLP [27]§	1.036	1.421	1.396	1.079	1.042	1.350	1.221	0.779
	DMD [11]§	1.019	1.380	1.367	0.976	0.984	1.278	1.167	0.746
	MCTN [18]‡	0.913	1.138	1.137	0.875	0.895	1.064	1.004	0.891
	TransM [19]‡	0.870	1.106	1.153	0.817	0.853	1.035	0.972	0.825
	SMIL [22]‡	0.894	1.067	1.112	0.866	0.859	1.019	0.970	0.818
	GCNet [16]‡	0.853	1.071	1.135	0.792	0.810	0.994	0.943	0.796
	CorrKD [32]‡	0.860	1.062	1.123	0.786	0.816	1.028	0.946	0.802
	HFGD (Ours)‡	**0.825**	**1.044**	**1.089**	**0.790**	**0.805**	**0.980**	**0.922***	0.768
MOSEI	Self-MM [7]§	0.723	1.308	1.367	0.701	0.717	1.278	1.016	0.548
	CubeMLP [27]§	0.768	1.353	1.428	0.725	0.750	1.301	1.054	0.540
	DMD [11]§	0.742	1.278	1.332	0.713	0.724	1.256	1.008	0.536
	MCTN [18]‡	0.654	1.125	1.138	0.668	0.654	1.080	0.887	0.607
	TransM [19]‡	0.661	1.107	1.160	0.630	0.651	1.102	0.885	0.577
	SMIL [22]‡	0.627	1.089	1.122	0.606	0.617	1.063	0.854	0.568
	GCNet [16]‡	0.602	1.064	1.107	0.586	0.597	1.016	0.829	0.545
	CorrKD [32]‡	0.610	1.058	1.097	0.575	0.589	1.007	0.823	0.556
	HFGD (Ours)‡	**0.587**	**1.032**	**1.067**	**0.565**	**0.572**	**0.987**	**0.802***	**0.533**

§ means the complete-modality methods and ‡ means the missing-modality methods. T-test is conducted on the "Avg." column, and * indicates that $p < 0.05$ (compared with SOTA CorrKD).

isons.

Experimental Setup. Regarding the MOSI [47] and MOSEI [48] datasets, we use the aligned multimodal sequences therein (*e.g.*, all sequences of modalities have length 300) as the original input for the HFGD. All models are implemented using the PyTorch framework and are trained on NVIDIA Tesla V100 GPUs. The Adam optimizer is used for optimizing the network. The detailed hyper-parameter settings for MOSI, MOSEI, and IEMOCAP are as follows: the learning rates are $\{4e-3, 4e-3, 2e-3\}$, the batch sizes are $\{32, 64, 64\}$, the epoch numbers are $\{50, 30, 50\}$, the attention heads are $\{10, 8, 10\}$, and the distance boundaries η are $\{1.2, 1.0, 0.7\}$. The embedding dimension is 40 on all three datasets. The hyper-parameters are determined using the validation set. The raw features at the modality missing positions are replaced by zero vectors. To guarantee an equitable comparison, we re-implement the State-Of-The-Art

Table 2: Comparison results under inter-modality missingness cases on MOSI and MOSEI. The F1 score metric is reported.

Datasets	Models	Testing Conditions							
		$\{l\}$	$\{a\}$	{v}	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	Avg.	$\{l,a,v\}$
MOSI	Self-MM [7]§	67.80	40.95	38.52	69.81	74.97	47.12	56.53	**84.64**
	CubeMLP [27]§	64.15	38.91	43.24	63.76	65.12	47.92	53.85	84.57
	DMD [11]§	68.97	43.33	42.26	70.51	68.45	50.47	57.33	84.50
	MCTN [18]‡	75.21	59.25	58.57	77.81	74.82	64.21	68.31	80.12
	TransM [19]‡	77.64	63.57	56.48	82.07	80.90	67.24	71.32	82.57
	SMIL [22]‡	78.26	67.69	59.67	79.82	79.15	71.24	72.64	82.85
	GCNet [16]‡	80.91	65.07	58.70	**84.73**	**83.58**	70.02	73.84	83.20
	CorrKD [32]‡	81.20	66.52	60.72	83.56	82.41	73.74	74.69	83.94
	HFGD (Ours)‡	**82.76**	**70.34**	**65.79**	84.38	83.39	**76.85**	**77.21***	84.30
MOSEI	Self-MM [7]§	71.53	43.57	37.61	75.91	74.62	49.52	58.79	83.69
	CubeMLP [27]§	67.52	39.54	32.58	71.69	70.06	48.54	54.99	83.17
	DMD [11]§	70.26	46.18	39.84	74.78	72.45	52.70	59.37	**84.78**
	MCTN [18]‡	75.50	62.72	59.46	76.64	77.13	64.84	69.38	81.75
	TransM [19]‡	77.98	63.68	58.67	80.46	78.61	62.24	70.27	81.48
	SMIL [22]‡	76.57	65.96	60.57	77.68	76.24	66.87	70.65	80.74
	GCNet [16]‡	80.52	66.54	61.83	81.96	81.15	69.21	73.54	82.35
	CorrKD [32]‡	80.76	66.09	62.30	81.74	81.28	71.92	74.02	82.16
	HFGD (Ours)‡	**82.34**	**69.46**	**66.09**	**83.25**	**82.68**	**75.56**	**76.56***	83.57

§ means the complete-modality methods and ‡ means the missing-modality methods. T-test is conducted on the "Avg." column, and * indicates that $p < 0.05$ (compared with SOTA CorrKD).

(SOTA) methods and integrate them with our experimental paradigms. All results are averages of multiple experiments under 10 different random seeds.

4.3. Comparison with State-of-the-art Methods

We select eight representative SOTA methods as baselines for comparison with HFGD. Specifically, in order to demonstrate the advantage of HFGD in the case of missing modality, we choose five missing-modality methods, including (1) joint learning methods, *i.e.*, MCTN [18], TransM [19], and CorrKD [32] and (2) generative methods *i.e.*, SMIL [22] and GCNet [16]. Additionally, to perform a more comprehensive comparison, we also select five complete-modality methods: Self-MM [7], CubeMLP [27], and DMD [11]. Thorough experiments were performed to comprehensively evaluate the robustness and applicability of the HFGD in two modality missing cases.

Table 3: Comparison results under inter-modality missingness cases on IEMOCAP. The F1 score metric is reported.

Models	Categories	Testing Conditions							
		$\{l\}$	$\{a\}$	$\{v\}$	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	Avg.	$\{l,a,v\}$
MulT [25]§	Happy	63.4	54.7	53.1	69.5	67.2	58.6	61.1	87.5
	Sad	64.2	53.2	51.9	71.3	69.7	57.2	61.3	85.3
	Angry	61.6	51.7	51.1	68.4	65.9	56.9	59.3	85.8
	Neutral	56.2	51.3	50.7	60.8	58.6	53.0	55.1	69.4
MICA [50]§	Happy	65.2	52.7	50.8	69.0	67.5	55.7	60.2	88.3
	Sad	64.8	52.9	52.5	67.4	65.4	56.3	59.9	86.6
	Angry	67.1	55.0	53.9	70.2	68.6	57.7	62.1	87.3
	Neutral	55.0	51.2	50.5	58.7	56.9	54.6	54.5	71.2
Self-MM [7]§	Happy	66.9	52.2	50.1	69.9	68.3	56.3	60.6	90.8
	Sad	68.7	51.9	54.8	71.3	69.5	57.5	62.3	86.7
	Angry	65.4	53.0	51.9	69.5	67.7	56.6	60.7	88.4
	Neutral	55.8	48.2	50.4	58.1	56.5	52.8	53.6	72.7
CubeMLP [27]§	Happy	68.9	54.3	51.4	72.1	69.8	60.6	62.9	89.0
	Sad	65.3	54.8	53.2	70.3	68.7	58.1	61.7	**88.5**
	Angry	65.8	53.1	50.4	69.5	69.0	54.8	60.4	87.2
	Neutral	53.5	50.8	48.7	57.3	54.5	51.8	52.8	71.8
DMD [11]§	Happy	69.5	55.4	51.9	73.2	70.3	61.3	63.6	**91.1**
	Sad	65.0	54.9	53.5	70.7	69.2	61.1	62.4	88.4
	Angry	64.8	53.7	51.2	70.8	69.9	57.2	61.3	**88.6**
	Neutral	54.0	51.2	48.0	56.9	55.6	53.4	53.2	72.2
MCTN [18]‡	Happy	76.9	63.4	60.8	79.6	77.6	66.9	70.9	83.1
	Sad	76.7	64.4	60.4	78.9	77.1	68.6	71.0	82.8
	Angry	77.1	61.0	56.7	81.6	80.4	58.9	69.3	84.6
	Neutral	60.1	51.9	50.4	64.7	62.4	54.9	57.4	67.7
TransM [19]‡	Happy	78.4	64.5	61.1	81.6	80.2	66.5	72.1	85.5
	Sad	79.5	63.2	58.9	82.4	80.5	64.4	71.5	84.0
	Angry	81.0	65.0	60.7	83.9	81.7	66.9	73.2	86.1
	Neutral	60.2	49.9	50.7	65.2	62.4	52.4	56.8	67.1
SMIL [22]‡	Happy	80.5	66.5	63.8	83.1	81.8	68.2	74.0	86.8
	Sad	78.9	65.2	62.2	82.4	79.6	68.2	72.8	85.2
	Angry	79.6	67.2	61.8	83.1	82.0	67.8	73.6	84.9
	Neutral	60.2	50.4	48.8	65.4	62.2	52.6	56.6	68.9
GCNet [16]‡	Happy	81.9	67.3	66.6	83.7	82.5	69.8	75.3	87.7
	Sad	80.5	69.4	66.1	83.8	81.9	70.4	75.4	86.9
	Angry	80.1	66.2	64.2	82.5	81.6	68.1	73.8	85.2
	Neutral	61.8	51.1	49.6	66.2	63.5	53.3	57.6	71.1
CorrKD [32]‡	Happy	82.6	69.6	68.0	84.1	82.0	70.0	76.1	87.5
	Sad	**82.7**	71.3	67.6	83.4	82.2	72.5	76.6	85.9
	Angry	82.2	67.0	65.8	**83.9**	82.8	67.3	74.8	86.1
	Neutral	63.1	54.2	52.3	68.5	64.3	57.2	59.9	71.5
HFGD (Ours)‡	Happy	**83.7**	**72.3**	**70.6**	**85.5**	**83.8**	**73.6**	**78.3***	89.6
	Sad	82.4	**73.5**	**71.1**	**84.7**	**82.2**	**74.6**	**78.1***	87.5
	Angry	**82.0**	**71.7**	**69.6**	83.2	**83.1**	**73.0**	**77.1***	87.2
	Neutral	**65.3**	**56.6**	**54.1**	**69.4**	**66.5**	**59.8**	**62.0***	**73.6**

§ means the complete-modality methods and ‡ means the missing-modality methods. T-test is conducted on the "Avg." column, and * indicates that $p < 0.05$ (compared with SOTA CorrKD).

Robustness to Intra-modality Missingness. We simulate intra-modality missingness by randomly discarding frame-level features in sequences with ratio $p \in \{0.1, 0.2, \cdots, 1.0\}$. The performance curves of models with different p values are illustrated in Figures 3 and 4, offering an intuitive depiction of the robustness of all models. Several key observations can be derived from these results. (i) As the ratio p increases, the performance of all models degrades. This observation indicates that intra-modality missingness results in substantial sentiment semantic loss and weakens the integrity of joint multimodal rep-

Table 4: Ablation results of components in HFGD on three datasets. "Intra-MM", "Inter-MM", and "CM" denote intra-modality missingness, inter-modality missingness, and complete modality.

Datasets	MOSI			MOSEI			IEMOCAP		
Metrics	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1
HFGD (Full)	**70.20**	**77.21**	**84.30**	**71.08**	**76.56**	**83.57**	**66.4**	**74.1**	**83.5**
w/o TSF	68.21	75.52	82.74	68.79	74.89	82.01	64.5	72.7	82.2
w/o DAF	68.46	75.16	83.23	69.41	75.02	82.69	65.0	73.1	82.7
w/o HSD	68.47	75.21	83.41	69.07	74.78	82.33	64.7	72.3	82.4

Table 5: Ablation experiments on MRM on three datasets. "Intra-MM", "Inter-MM", and "CM" denote intra-modality missingness, inter-modality missingness, and complete modality.

Datasets	MOSI			MOSEI			IEMOCAP		
Metrics	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1
HFGD (Full)	**70.20**	**77.21**	**84.30**	**70.20**	**77.21**	**84.30**	**70.20**	**77.21**	**84.30**
w/o intra-MRM	67.95	76.06	84.24	67.95	76.06	84.24	67.95	76.06	84.24
w/o inter-MRM	69.67	74.80	84.46	69.67	74.80	84.46	69.67	74.80	84.46

Table 6: Ablation experiments of the TSF on three datasets. "Intra-MM", "Inter-MM", and "CM" denote intra-modality missingness, inter-modality missingness, and complete modality.

Datasets	MOSI			MOSEI			IEMOCAP		
Metrics	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1
HFGD (Full)	**70.20**	**77.21**	**84.30**	**71.08**	**76.56**	**83.57**	**66.4**	**74.1**	**83.5**
w/o ω	69.65	76.71	83.95	70.23	76.01	82.89	65.8	73.5	82.7
w/o $\mathcal{L}_{TSF}^{intra}$	68.92	76.20	83.79	69.77	75.67	82.47	65.1	73.2	82.3
w/o $\mathcal{L}_{TSF}^{inter}$	68.45	75.87	83.26	69.24	75.35	82.14	64.8	72.9	81.9

Table 7: Ablation experiments of DAF on three datasets. "Intra-MM", "Inter-MM", and "CM" denote intra-modality missingness, inter-modality missingness, and complete modality.

Datasets	MOSI			MOSEI			IEMOCAP		
Metrics	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1
HFGD (Full)	**70.20**	**77.21**	**84.30**	**71.08**	**76.56**	**83.57**	**66.4**	**74.1**	**83.5**
w/o $\mathcal{L}_{DAF}^{crtst}$	69.25	76.03	83.75	70.32	75.89	83.23	65.2	73.4	83.1
w/o $\mathcal{L}_{DAF}^{adv}$	68.82	75.46	83.40	69.78	75.44	82.97	65.6	73.7	83.4

Table 8: Ablation experiments HSD on three datasets. "Intra-MM", "Inter-MM", and "CM" denote intra-modality missingness, inter-modality missingness, and complete modality.

Datasets	MOSI			MOSEI			IEMOCAP		
Metrics	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1
HFGD (Full)	**70.20**	**77.21**	**84.30**	**71.08**	**76.56**	**83.57**	**66.4**	**74.1**	**83.5**
w/o MF	69.62	76.82	84.19	70.82	76.19	83.24	66.2	73.7	83.2
w/o $\mathcal{L}_{HSD}^{feat}$	68.79	75.63	83.56	69.45	75.04	82.69	65.0	72.8	82.6
w/o $\mathcal{L}_{HSD}^{logits}$	69.23	76.11	83.92	70.42	75.69	82.90	65.4	73.4	83.0

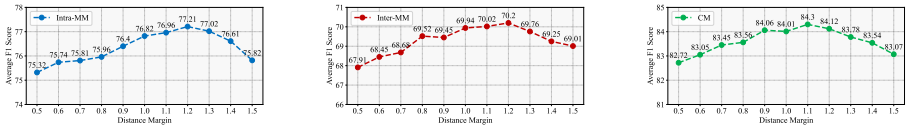
resentations. (ii) Compared to complete-modality methods, our HFGD exhibits significant performance benefits under missing-modality testing conditions and shows competitive results with complete modalities. This advantage comes from the proposed two-stage modality factorization strategy for mining high-quality sentiment semantics and the hierarchical knowledge distillation module for precise recovery of missing sentiment information. This enables HFGD to exhibit superior performance in both complete and missing modality cases, resulting in good generalization and applicability in real-world applications. (iii) Contrary to missing-modality methods, our HFGD exhibits superior robustness. Through the factorization of modalities and the hierarchical reconstruction paradigm in the self-distillation module, HFGD effectively restores missing features and generates robust multimodal representations. (iv) An intriguing phenomenon is that when all features within modalities are missing (*i.e.*, $p = 1.0$), the results converge to a smaller interval on MOSI and MOSEI. We attribute this finding to label bias, representing spurious statistical shortcuts captured by MSA models after training on datasets with unbalanced sample distributions. When no features are available in the modalities, the models' predictions would converge to the corresponding ideal distribution of the label bias.

Robustness to Inter-modality Missingness. To simulate testing conditions of inter-modality missingness, we remove certain entire modalities from the samples. Tables 2 and 3 contrast the models' resilience to inter-modality missingness. In these comparisons, "$\{l\}$" indicates that only the language modality is

available, with the audio and visual modalities missing. "$\{l, a, v\}$" represents the complete-modality testing condition where all modalities are available. "Avg." refers to the average performance across six different missing-modality testing conditions. We have the following key findings: (i) The cases of inter-modality missingness lead to a decline in performance for all models, indicating that integrating complementary information from diverse modalities enhances the sentiment semantics within joint representations. (ii) Under the modality missing testing conditions, the performance degradation of the complete-modality methods is much larger than that of the missing-modality methods, suggesting that the missing-modality methods utilize their semantic-recovery learning paradigm to mitigate the disaster of modality missing to some extent. Among all the methods, our HFGD has the best performance, proving its strong robustness. This advantage stems from its sufficient capture of sentiment semantics based on temporal saliency and distributional alignment-based factorization and precise recovery and reconstruction of missing features via the hierarchical self-distillation mechanism. The advantage comes from its learning of adequate representation factorization and hierarchical feature consistency supervision. (iii) Distinct modalities have varying importance in sentiment understanding. Specifically, in the unimodal testing conditions, all methods perform best when using only language modality, and have comparable performance to the complete modality case. In the bimodal testing condition, combinations containing language modality show better performance, and our HFGD even outperforms the complete modality case in individual metrics. This phenomenon proves that language modality contains the richest sentiment semantics and plays a dominant role in sentiment inference and missing semantics reconstruction.

4.4. Ablation Studies

Ablation of Components in HFGD. To validate the effectiveness and necessity of the proposed components in HFGD, we conducted comprehensive ablation experiments on the two missing-modality cases of the three datasets, as shown in Table 4. The testing conditions consist of Intra-Modality Missingness

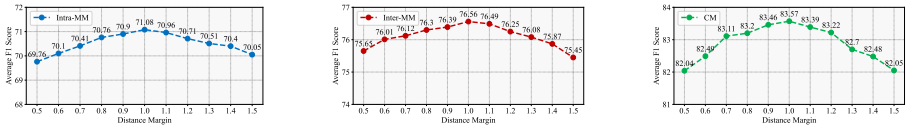


(a) Average F1 score for cases of Intra-modality missingness.

(b) Average F1 score for cases of inter-modality missingness.

(c) F1 Score for cases of complete modality.

Figure 5: Performance of HFGD with different distance margins in DAF on the MOSI.



(a) Average F1 Score for cases of intra-modality missingness.

(b) Average F1 Score for cases of inter-modality missingness.

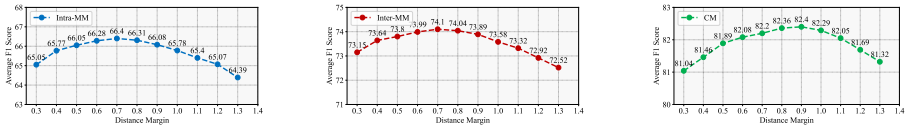
(c) F1 Score for cases of complete modality.

Figure 6: Performance of HFGD with different distance margins in DAF on the MOSEI.

(Intra-MM), Intra-Modality Missingness (Inter-MM), and Complete Modality (CM). We only report the F1 score for brevity. (i) Firstly, we removed the TSF, and the consistent performance degradation in both missing-modality cases demonstrates that the factorization mechanism based on temporal saliency is able to capture critical information and sentiment semantics in modalities by exploiting the saliency between elements on multimodal sequences.

(ii) Then, when our DAF is eliminated, the worse performance demonstrates that adequately aligning the distribution between representations leveraging contrastive and adversarial learning can effectively facilitate the recovery of missing sentiment semantics and the generation of robust joint multimodal representations. (iii) Finally, we remove HSD, and the declined results illustrate that the hierarchical self-distillation paradigm effectively learns consistency between heterogeneous representations, thus incrementally mining and capturing missing semantics.

Ablation Results of MRM. Table 5 presents ablation studies of the proposed Modality Random Missing (MRM) strategy to verify its necessity. MRM includes Intra-modality MRM (Intra-MRM) and Inter-modality MRM (Inter-MRM), where the features in the missing part are replaced with zero vectors.



(a) Average F1 Score for cases of intra-modality missingness.

(b) Average F1 Score for cases of inter-modality missingness.

(c) F1 Score for cases of complete modality.

Figure 7: Performance of HFGD with different distance margins in DAF on the IEMOCAP.

Concretely, we remove Intra-MRM and Inter-MRM in the training phase to observe the model performance changes under different testing conditions on the MOSI, MOSEI, and IEMOCAP, respectively. When different cases of modality missingness are excluded, the corresponding test conditions show consistent performance degradation. This phenomenon demonstrates the necessity of implementing frame-level and modality-level missing operations. The reason is that implementing both intra- and inter-MRM in the training phase can maximize the potential of the hierarchical self-distillation module in HFGD to recover the missing semantics, thus realistically simulating sophisticated modality missing situations in real-world applications.

Ablation Results of TSF. To validate the necessity of ω, $\mathcal{L}_{TSF}^{intra}$, and $\mathcal{L}_{TSF}^{inter}$ in TSF, we conducted ablation experiments on the MOSI, MOSEI, and IEMOCAP datasets, which are shown in Table 6. We have the following important findings: (i) when the temporal saliency weight ω is removed, and all segments sliced within the modality sequence are assigned the same weight during the learning process. The degraded performance demonstrates that the model cannot focus on the most sentimentally rich local features in the modality sequences and cannot capture valuable information effectively. (ii) We removed $\mathcal{L}_{TSF}^{intra}$, and the significantly reduced performance indicates that the factorization process is facilitated by maximizing the mutual information between segments with different significance and the factorized representations, thus ensuring the precise reconstruction of the missing semantics. (iii) The poorer performance when $\mathcal{L}_{TSF}^{inter}$ is removed suggests that enforcing dynamic consistency constraints between heterogeneous representations across modalities facilitates capturing global and

holistic sentiment information of samples. Overall, the consistently decreased results in all testing conditions on the three datasets confirm the reasonableness of our default choice.

Ablation Results of DAF. In Table 7, we perform thorough ablation studies of the distributional alignment learning paradigm in DAF. Extensive experiments are conducted on the MOSI, MOSEI, and IEMOCAP. We remove the contrastive learning and adversarial learning paradigms, respectively, and the dramatic performance degradation indicates that considering only semantic-level factorization and ignoring distribution-level factorization cannot effectively extract sentiment information. When the contrastive learning loss $\mathcal{L}_{DAF}^{ctrst}$ is removed, the model performance decreases significantly. The contrastive learning paradigm learns distributional consistency between collaborative representations as well as the distributional dissimilarity between collaborative representation and modality-exclusive representations. The core idea is to decouple heterogeneous representations in the latent feature space and enhance the integrity and independence of their distributions. When the adversarial learning loss $\mathcal{L}_{DAF}^{adv}$ is removed, the worse performance proves that adversarial learning facilitates and normalizes the distribution between heterogeneous representations, thus sufficiently factorizing and decoupling representations with different semantics for precise reconstruction of subsequent missing features. In summary, contrastive and adversarial learning paradigms contribute to the recovery of missing semantics and the generation of robust multimodal representations, which assume an indispensable role in the framework.

Moreover, we further evaluate the effect of the predefined distance margin η in contrastive learning of DAF on performance. Figures 5, 6, and 7 show the performance variations for different values of the distance margin on the MOSI, MOSEI, and IEMOCAP, respectively. The performance of the model varies with the boundary η, suggesting that defining an appropriate boundary is important for mining the differences between two objects in negative pairs in contrastive learning. The optimal values of distance margin η on the three datasets are 1.2, 1.0, and 0.7, respectively. Intuitively, selecting appropriate margin values

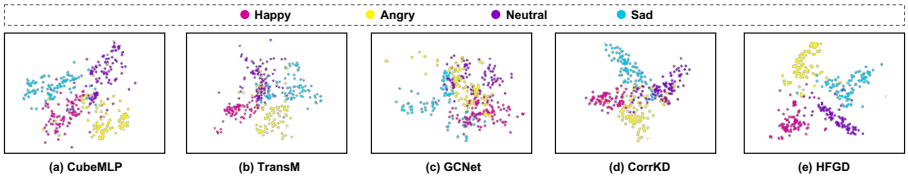


Figure 8: Representation visualization of partial samples on the IEMOCAP dataset. The testing conditions include intra-modality missingness (*i.e.*, missing ratio $p = 0.5$) and inter-modality missingness (*i.e.*, only the language modality is available).

on different datasets favors the network to learn more discriminative sentiment semantics, leading to better gains.

Ablation Results of HSD. To evaluate the effectiveness of the components in HSD, we perform comprehensive ablation studies on the MOSI, MOSEI, and IEMOCAP in Table 8. Obviously, the full HFGD has a better performance than the other variants under all testing conditions. Specifically, (i) we remove the Multi-scale Features (MF), and the performance decreased noticeably. This is because the self-distillation framework utilizes MF to ensure hierarchical consistency between the heterogeneous representations of both networks, thus hierarchically transferring more detailed and comprehensive knowledge and facilitating the integrated recovery of missing semantics. (ii) The Wasserstein distance-based feature distillation loss is substituted for the KL divergence, and the reduced evaluation metrics indicate that the Wasserstein distance effectively computes and constrains the distributional difference between the two heterogeneous representations based on OMT problem, which effectively ensures the feature-level consistency. (iii) Consistency supervision at the logits level ensures that two heterogeneous representations maintain similar probability distributions and decision boundaries.

4.5. Qualitative Analysis

In order to further demonstrate the performance advantage of HFGD under modality missing, we randomly selected 100 samples from each emotion category in the IEMOCAP testing set for visual evaluation. The compari-

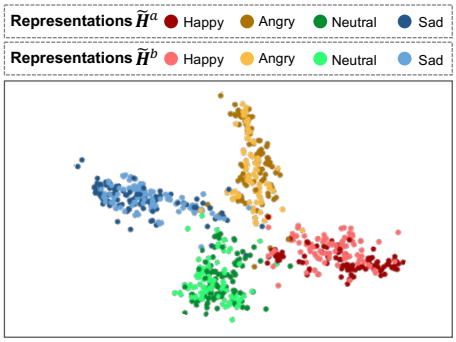


Figure 9: Visualization of $\tilde{\boldsymbol{H}}^a$ and $\tilde{\boldsymbol{H}}^b$ of HFGD in IEMOCAP.

son models include CubeMLP [27] (complete-modality method), TransM [19] (joint learning-based missing-modality method), GCNet [16] (generation-based missing-modality method), and CorrKD [32] (joint learning-based missing-modality method). (i) As shown in Figure 8, CubeMLP faces significant difficulties when dealing with missing modalities, as the representations of different emotion categories become heavily entangled, resulting in substantially poor performance. (ii) While TransM, GCNet, and CorrKD alleviate some of the indistinguishable emotion semantics, their performance remains sub-optimal due to the generally ambiguous and overlapping distribution boundaries of different emotion representations. (iii) Conversely, our HFGD facilitates the formation of dense clusters for representations within the same emotion category and distinctly separates those from different categories.

In Figure 9, we visualize the representations $\tilde{\boldsymbol{H}}_a$ and $\tilde{\boldsymbol{H}}_b$ on the IEMOCAP dataset. 100 samples in each emotion category are randomly selected, where dots of the same color scheme represent feature distribution to the same emotion category from both representations. We have the following important findings:(i) $\tilde{\boldsymbol{H}}_a$ and $\tilde{\boldsymbol{H}}_b$ have highly similar distributions in the feature space. This proves that the proposed hierarchical knowledge distillation paradigm can sufficiently deliver beneficial knowledge and motivate the student network to recover the missing features and sentiment semantics. (ii) For the student network, the different sentiment categories are well grouped and form compact

clusters, demonstrating that our HFGD is highly robust to the modality missing cases. These observations confirm that the proposed framework ensures that two heterogeneous joint multimodal representations can capture consistent semantics and knowledge, leading to more robust sentiment predictions.

5. Conclusion and Discussion

In this paper, we present a Hierarchical Factorization-Guided Self-Distillation (HFGD) framework to address diverse missing modality dilemmas in the MSA task. Specifically, we propose a temporal saliency-based factorization mechanism that dynamically maximizes the mutual information between the representations based on temporal saliency to refine the sentiment semantics. Moreover, we propose a distributional alignment-based factorization mechanism that further aligns the representation distributions utilizing both contrastive and adversarial learning. Eventually, a hierarchical self-distillation module is proposed to precisely reconstruct the missing semantics by multi-scale consistency supervision. Thorough experiments validate the indispensability of our framework.

Discussion of Broad Impacts. The positive impact of our approach lies in the ability to significantly improve the robustness and stability of multimodal sentiment analysis systems against heterogeneous modality missingness in real-world applications. Nevertheless, this technology may have a negative impact when it falls into the wrong hands, e.g., the proposed model is used for malicious purposes by injecting biased priors to recognize the emotions of specific groups.

Discussion of Limitation and Future Work. The current method defines the modality missing cases as both inter-modality missingness and intra-modality missingness. Nevertheless, in real-world applications, modality missing cases may be very intricate and difficult to simulate. Consequently, the proposed method may suffer some minor performance loss when applied to real-world scenarios. In the future, we will explore more intricate modality missing cases and design suitable algorithms to compensate for this deficiency.

References

[1] Y. Du, D. Yang, P. Zhai, M. Li, L. Zhang, Learning associative representation for facial expression recognition, in: Proc. Int. Conf. Image Process., 2021, pp. 889–893.

[2] D. Yang, K. Yang, M. Li, S. Wang, S. Wang, L. Zhang, Robust emotion recognition in context debiasing, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2024.

[3] Z. Xie, Y. Yang, J. Wang, X. Liu, X. Li, Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space, IEEE Trans. Circuits Syst. Video Technol.

[4] M. Li, J. Shi, L. Bai, C. Huang, Y. Jiang, K. Lu, S. Wang, E. R. Hancock, Frameerc: Framelet transform based multimodal graph neural networks for emotion recognition in conversation, Pattern Recognit. 161 (2025) 111340.

[5] Z. Liu, L. Cai, W. Yang, J. Liu, Sentiment analysis based on text information enhancement and multimodal feature fusion, Pattern Recognit. 156 (2024) 110847.

[6] G. Xiang, S. Yao, X. Wu, H. Deng, G. Wang, Y. Liu, F. Li, Y. Peng, Driver multi-task emotion recognition network based on multi-modal facial video analysis, Pattern Recognit. 161 (2025) 111241.

[7] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proc. AAAI Conf. Artif. Intell., Vol. 35, 2021, pp. 10790–10797.

[8] D. Yang, S. Huang, H. Kuang, Y. Du, L. Zhang, Disentangled representation learning for multimodal emotion recognition, in: Proc. ACM Int. Conf. Multimedia, 2022, pp. 1642–1651.

[9] D. Yang, H. Kuang, S. Huang, L. Zhang, Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences, in: Proc. ACM Int. Conf. Multimedia, 2022, pp. 1708–1717.

[10] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, L. Zhang, Emotion recognition for multiple context awareness, in: Proc. Eur. Conf. Comput. Vis., Springer, 2022, pp. 144–162.

[11] Y. Li, Y. Wang, Z. Cui, Decoupled multimodal distilling for emotion recognition, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 6631–6640.

[12] D. Yang, M. Li, L. Qu, K. Yang, P. Zhai, S. Wang, L. Zhang, Asynchronous multimodal video sequence fusion via learning modality-exclusive and-agnostic representations, IEEE Trans. Circuits Syst. Video Technol.

[13] W. Zou, X. Sun, Q. Lu, X. Wang, J. Feng, A vision and language hierarchical alignment for multimodal aspect-based sentiment analysis, Pattern Recognit. (2025) 111369.

[14] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259.

[15] 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: Proc. ACM Int. Conf. Multimedia, 2018, pp. 108–116.

[16] Z. Lian, L. Chen, L. Sun, B. Liu, J. Tao, Gcnet: graph completion network for incomplete multimodal learning in conversation, IEEE Trans. Pattern Anal. Mach. Intell.

[17] Y. Wang, Z. Cui, Y. Li, Distribution-consistent modal recovering for incomplete multimodal learning, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 22025–22034.

[18] 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: Proc. AAAI Conf. Artif. Intell., Vol. 33, 2019, pp. 6892–6899.

[19] Z. Wang, Z. Wan, X. Wan, Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis, in: Proc. Web Conf., 2020, pp. 2514–2520.

[20] J. Zeng, T. Liu, J. Zhou, Tag-assisted multimodal sentiment analysis under uncertain missing modalities, in: Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2022, pp. 1545–1554.

[21] M. Li, D. Yang, Y. Lei, S. Wang, S. Wang, L. Su, K. Yang, Y. Wang, M. Sun, L. Zhang, A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities, in: Proc. AAAI Conf. Artif. Intell., Vol. 38, 2024, pp. 10074–10082.

[22] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, X. Peng, Smil: Multimodal learning with severely missing modality, in: Proc. AAAI Conf. Artif. Intell., Vol. 35, 2021, pp. 2302–2310.

[23] X.-S. Wei, H.-T. Yu, A. Xu, F. Zhang, Y. Peng, Mecom: A meta-completion network for fine-grained recognition with incomplete multi-modalities, IEEE Trans. Image Process.

[24] T. Qin, B. Peng, J. Lei, J. Song, L. Xu, Q. Huang, 3d-immc: Incomplete multi-modal 3d shape clustering via cross mapping and dual adaptive fusion, IEEE Trans. Emerging Top. Comput. Intell.

[25] 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: Proc Conf. Assoc. Comput. Linguist. Meet., Vol. 2019, NIH Public Access, 2019, p. 6558.

[26] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proc. Conf. Empir. Methods Nat. Lang. Process.,, 2021, pp. 9180–9192.

[27] H. Sun, H. Wang, J. Liu, Y.-W. Chen, L. Lin, Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation, in: Proc. ACM Int. Conf. Multimedia, 2022, pp. 3722–3729.

[28] R. Chen, W. Zhou, Y. Li, H. Zhou, Video-based cross-modal auxiliary network for multimodal sentiment analysis, IEEE Trans. Circuits Syst. Video Technol. 32 (12) (2022) 8703–8716.

[29] Z. Guo, T. Jin, Z. Zhao, Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition, arXiv preprint arXiv:2407.05374.

[30] W. Xu, H. Jiang, X. Liang, Leveraging knowledge of modality experts for incomplete multimodal learning, in: Proc. ACM Int. Conf. Multimedia, 2024, pp. 438–446.

[31] Z. Gao, D. Hu, X. Jiang, H. Lu, H. T. Shen, X. Xu, Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis, in: Proc. ACM Int. Conf. Multimedia, 2024, pp. 9650–9659.

[32] M. Li, D. Yang, X. Zhao, S. Wang, Y. Wang, K. Yang, M. Sun, D. Kou, Z. Qian, L. Zhang, Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2024, pp. 12458–12468.

[33] H. Kim, A. Mnih, Disentangling by factorising, in: Proc. Int. Conf. Mach. Learn., PMLR, 2018, pp. 2649–2658.

[34] X. Wu, H. Huang, V. M. Patel, R. He, Z. Sun, Disentangled variational representation for heterogeneous face recognition, in: Proc. AAAI Conf. Artif. Intell., Vol. 33, 2019, pp. 9005–9012.

[35] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3967–3976.

[36] J. H. Cho, B. Hariharan, On the efficacy of knowledge distillation, in: Proc. IEEE/CVF Int. Conf. Comput. Vision, 2019, pp. 4794–4802.

[37] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, H. Ghasemzadeh, Improved knowledge distillation via teacher assistant, in: Proc. AAAI Conf. Artif. Intell., Vol. 34, 2020, pp. 5191–5198.

[38] J. W. Cho, D.-J. Kim, J. Choi, Y. Jung, I. S. Kweon, Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1592–1601.

[39] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, P. Gori, Knowledge distillation from multi-modal to mono-modal segmentation networks, in: Proc. Conf. Med. Image Comput. Comput. Assisted Intervention, Springer, 2020, pp. 772–781.

[40] H. Wang, C. Ma, J. Zhang, Y. Zhang, J. Avery, L. Hull, G. Carneiro, Learnable cross-modal knowledge distillation for multi-modal learning with missing modality, in: Proc. Conf. Med. Image Comput. Comput. Assisted Intervention, Springer, 2023, pp. 216–226.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30.

[42] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, in: Int. Conf. on Learn. Representations, 2018.

[43] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, M. Botvinick, On the importance of single directions for generalization, Stat 1050 (2018) 15.

[44] C. Villani, C. Villani, The wasserstein distances, Optim. Transport Old New (2009) 93–111.

[45] C. Villani, et al., Optim. Transport Old New, Vol. 338, Springer, 2009.

[46] S. Kolouri, G. K. Rohde, H. Hoffmann, Sliced wasserstein distance for learning gaussian mixture models, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3427–3436.

[47] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv preprint arXiv:1606.06259.

[48] 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: Proc. Annu. Meet. Assoc. Comput. Ling., 2018, pp. 2236–2246.

[49] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.

[50] T. Liang, G. Lin, L. Feng, Y. Zhang, F. Lv, Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion, in: Proc. IEEE/CVF Int. Conf. Comput. Vision, 2021, pp. 8148–8156.

[51] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proc. Conf. Empirical Methods Nat. Lang. Process., 2014, pp. 1532–1543.

[52] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep—a collaborative voice analysis repository for speech technologies, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Process., IEEE, 2014, pp. 960–964.

[53] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: Proc. IEEE Winter Conf. Appl. Comput. Vis., 2016, pp. 1–10.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:
