FSRF: Factorization-guided Semantic Recovery for Incomplete Multimodal Sentiment Analysis

Anonymous ICME submission

Abstract—**In recent years, Multimodal Sentiment Analysis (MSA) has become a research hotspot that aims to utilize multimodal data for human sentiment understanding. Previous MSA studies have mainly focused on performing interaction and fusion on complete multimodal data, ignoring the problem of missing modalities in real-world applications due to occlusion, personal privacy constraints, and device malfunctions, resulting in low generalizability. To this end, we propose a Factorization-guided Semantic Recovery Framework (FSRF) to mitigate the modality missing problem in the MSA task. Specifically, we propose a de-redundant homo-heterogeneous factorization module that factorizes modality into modality-homogeneous, modality-heterogeneous, and noisy representations and design elaborate constraint paradigms for representation learning. Furthermore, we design a distribution-aligned self-distillation module that fully recovers the missing semantics by utilizing bidirectional knowledge transfer. Comprehensive experiments on two datasets indicate that FSRF has a significant performance advantage over previous methods with uncertain missing modalities.**

Index Terms—**multimodal sentiment analysis, modality missing, representation learning**

I. Introduction

Multimodal Sentiment Analysis (MSA) has become a prominent focus in recent years. Unlike conventional emotion recognition tasks that rely on a single modality [1], MSA integrates various modalities to enhance the understanding of human sentiment [2]. Prior research has demonstrated that integrating complementary semantics across multiple modalities enhances the creation of more precise multimodal representations [3]. MSA has been well studied so far under the assumption that all modalities are available in the training and inference phases [4]–[8]. However, in practical applications, modalities are often missing due to factors such as security issues, background noise, or sensor limitations. These incomplete multimodal datasets substantially impair the performance of MSA.

In recent years, many studies [9]–[15] have focused on tackling the challenge of missing modalities in MSA, which can be classified into two distinct paradigms: (i) generative methods [9], [10] and (ii) joint learning methods [11], [13], [15]. Generative methods focus on reconstructing missing features and semantics within modalities by utilizing the distributions of the accessible modalities. For example, GCNet [9] proposes two graph neural network-based modules to capture speaker and temporal dependencies in conversations to address the problem of missing modality in MSA. In contrast, joint learning methods aim to develop joint multimodal representations by exploiting correlations between modalities. For instance, CorrKD [15] introduces a correlation-decoupled knowledge distillation framework that leverages cross-sample, cross-category, and cross-response correlations to create robust joint multimodal representations for mitigating the modality missing problem in MSA. However, these approaches are limited by two factors: (1) Implementing complex inter-modality interactions based on incomplete modalities ignores the extraction of task-relevant semantics and task-independent noise. (2) The recovery paradigm for missing semantics is coarse-grained and static, without sufficient consideration of high-dimensional distributional alignment.

To address the above problem, we propose a Factorization-guided Semantic Recovery Framework (FSRF) to solve the modality missing problem in MSA. Our strength comes from the following three core contributions: (i) We propose a de-redundant homo-heterogeneous factorization module that decomposes each modality into modality-homogeneous, modality-heterogeneous, and noise representations, and design elaborate constraint paradigms for representation learning. (ii) We propose a distribution alignment-based self-distillation module that utilizes the Sinkhorn distance and JS divergence to achieve bidirectional constraints between two networks. (iii) Extensive experiments conducted on two datasets demonstrate that FSRF has a significant performance advantage over previous methods with uncertain missing modalities and comparable performance with complete modalities.

II. Methodology

A. Problem Formulation

Given a multimodal video dataset $\mathcal{D} = \{\boldsymbol{X}_i, \boldsymbol{Y}_i\}_{i=1}^N$, where N is the number of samples. Each sample is containing three distinct modalities, denoted as $\boldsymbol{S} = [\boldsymbol{M}_L, \boldsymbol{M}_A, \boldsymbol{M}_V]$, where $\boldsymbol{M}_L \in \mathbb{R}^{T_L \times d_L}, \boldsymbol{M}_A \in \mathbb{R}^{T_A \times d_A}$, and $\boldsymbol{M}_V \in \mathbb{R}^{T_V \times d_V}$ denote language, audio, and visual modalities, respectively. $\Phi = \{L, A, V\}$ denotes the set of modality types. $T_m(\cdot)$ and $d_m(\cdot)$ represent the sequence length and the embedding dimension, where $m \in \Phi$. To effectively replicate the uncertainty of modality missingness observed in real-world scenarios, we define two types of 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. The goal is to perform utterance-level sentiment recognition by utilizing multimodal data that includes missing modalities.

B. Modality Preprocessing

As shown in Figure 1, the Modality Random Missing (MRM) strategy performs both intra-modality missing and inter-modality missing, generating two heterogeneous modality

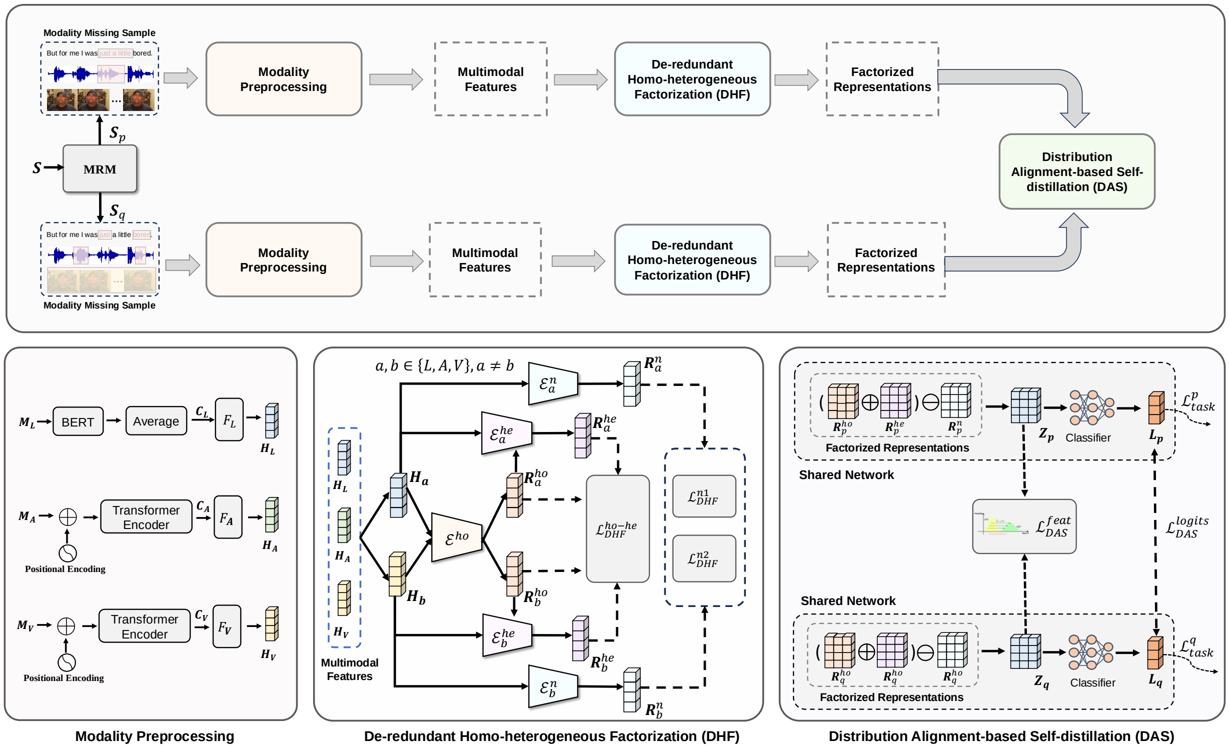

Fig. 1: The overall architecture of FSRF. We propose a De-redundant Homo-Heterogeneous Factorization (DHF) mechanism and a Distributional Alignment-based Self-Distillation (DAS) module to accurately recover missing information.

missing samples $\boldsymbol{S}_p$ and $\boldsymbol{S}_q$ from the input complete modality sample $\boldsymbol{S}$. Our framework receives $\boldsymbol{S}_p$ and $\boldsymbol{S}_q$ as input in parallel and symmetrically. For brevity, here we denote a particular sample of the input to the network with $\boldsymbol{S} = [\boldsymbol{M}_L, \boldsymbol{M}_A, \boldsymbol{M}_V]$. The language modality is fed into the pre-trained BERT [16] to obtain the average feature for each word, denoted as $\boldsymbol{C}_L \in \mathbb{R}^{d_L}$. The audio and visual modalities plus the positional encoding are then fed into the Transformer encoder [17], whose last elements of the output are expressed as $\boldsymbol{C}_A \in \mathbb{R}^{d_A}$ and $\boldsymbol{C}_V \in \mathbb{R}^{d_V}$. Each $\boldsymbol{C}_m$ is passed through a fully-connected layer to unify the dimensions of the features to obtain $\boldsymbol{H}_m$.

C. De-redundant Homo-Heterogeneous Factorization

Modality missing destroys the contextual semantics inherent in multimodal data, significantly impacting the effectiveness of multimodal fusion and semantic extraction. Previous studies of MSA with missing modalities [11], [12], [18] focus on designing complex cross-modal interaction paradigms to mine valuable information in incomplete modalities. These approaches treat incomplete modalities as the minimal interaction unit and ignore the shared, specific, and noisy information contained in the modalities, thus introducing significant cumulative errors. Therefore, we propose a De-redundant Homo-heterogeneous Factorization (DHF) module, to fully exploit the multilevel representations in modality.

Specifically, the DHF factorizes each modality into three categories of representations: (1) Modality-homogeneous representation, denoting the representation reflecting the sentiment homogeneity shared among all modalities. (2) Modality-heterogeneous representation, which represents different modality-specific information for characterizing sentiment in different modalities. (3) Noise representation, which indicates the noise information contained in each modality. For each modality $m \in \Phi$, we obtain the modality-homogeneous representation by a modality-shared encoder, denoted as $\boldsymbol{R}_m^{ho} = \mathcal{E}^{ho}(\boldsymbol{H_m})$, a modality-heterogeneous representation by a modality-specific encoder, denoted as $\boldsymbol{R}_m^{he} = \mathcal{E}_m^{he}(\boldsymbol{H}_m, \boldsymbol{R}_m^{ho})$, and a noise representation by a modality-specific encoder, denoted as $\boldsymbol{R}_m^n = \mathcal{E}_m^n(\boldsymbol{H}_m)$. All encoders are composed of multi-layer perceptrons with the ReLU activation.

Modality-homogeneous representations should contain sentiment information shared among all modalities in the sample, and modality-heterogeneous representations of different modalities should be inconsistent. To achieve this objective, we propose a contrastive learning-based representation constraint mechanism, which employs NT-Xent loss [19] to constrain the distribution of different representations in the latent space. It has stronger constraints and training stability than traditional contrastive learning loss through normalization and temperature parameters. The loss is denoted as:

$$\mathcal{L}_{DHF}^{ho-he} = -\frac{1}{6}\sum_{a\in\Phi}\sum_{b\in\Phi,a\neq b} log\frac{\mathcal{G}(\boldsymbol{R}_a^{ho},\boldsymbol{R}_b^{ho})}{\mathcal{G}(\boldsymbol{R}_a^{ho},\boldsymbol{R}_b^{ho})+\mathcal{G}(\boldsymbol{R}_a^{ho},\boldsymbol{R}_a^{he})}, \tag{1}$$

where $\mathcal{G}(\boldsymbol{X}, \boldsymbol{Y}) = exp(-\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y})/\tau)$, $\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y})$ is the cosine distance of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and τ is the temperature hyperparameter. Additionally, each modality has its unique noise representation while remaining consistent across samples of the same modality, which is constrained as follows:

$$\mathcal{L}^{n_1}_{DHF} = \frac{1}{3\,N_B\,(N_B-1)}\sum_{i=1}^{N_B}\sum_{j=1,j\neq i}^{N_B}\sum_{m\in\Phi}\mathcal{D}\left(\boldsymbol{R}^n_{i,m},\boldsymbol{R}^n_{j,m}\right) + \frac{1}{6\,N_B}\sum_{i=1}^{N_B}\sum_{a\in\Phi}\sum_{b\in\Phi,b\neq a}\max\left(\epsilon-\mathcal{D}\left(\boldsymbol{R}^n_{i,a},\boldsymbol{R}^n_{i,b}\right),0\right), \tag{2}$$

where N_B is the number of samples in a mini-batch and ϵ is the predefined distance margin. Moreover, to enhance the compactness of the noise representation in low-dimensional space to facilitate the stable extraction of noisy information, we minimize the entropy of the noise representation. Simultaneously, we use regularization to control the noise representation to avoid it being excessively compact and losing diversity. This regularized entropy loss is expressed as:

$$\mathcal{L}^{n_2}_{DHF} = \frac{1}{3}\sum_{m\in\Phi}\left(\mathcal{H}\left(\boldsymbol{R}^n_m\right) + (\|\boldsymbol{R}^n_m - \mu_m\|^2 + \left\|Var(\boldsymbol{R}^n_m) - \sigma^2_m)\right\|^2\right), \tag{3}$$

where $\mathcal{H}(\cdot) = \frac{1}{2}\log\left((2\pi e)^d\det(\Sigma)\right)$ is the entropy formula for the multi-dimensional Gaussian distribution, and Σ is the covariance matrix, and μ_m and σ^2_m are the predefined mean and variance. Therefore, the loss of DHF is expressed as:

$$\mathcal{L}_{DHF} = \mathcal{L}^{ho-he}_{DHF} + \mathcal{L}^{n_1}_{DHF} + \mathcal{L}^{n_2}_{DHF}. \tag{4}$$

D. Distribution Alignment-based Self-Distillation

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 [20], [21]. To tackle the aforementioned challenges, we propose a Distribution Alignment-based Self-distillation (DAS) module to progressively learn the representation consistency and recover the missing sentiment semantics through a hierarchical self-distillation paradigm. Specifically, DAS performs bi-directional information delivery within a unified and shared network to achieve feature and logits consistency constraints between two heterogeneous modality missing samples. This learning paradigm reduces the unidirectional dependence on knowledge [22] and provides two key advantages: transferring knowledge from richer modalities to those with fewer modalities helps recover information lost due to missing modalities, while the reverse flow of knowledge refines and strengthens useful features. In conclusion, DAS enables the model to generate more robust joint multimodal representations.

Feature-level Distillation. We add all the modality-homogeneous representations and modality-heterogeneous representations and subtract all the noisy representations to obtain the joint multimodal representations $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_q$ of the two heterogeneous samples. To accurately measure the holistic discrepancy between the distributions of the two representations, we utilize the Sinkhorn distance [23]. It is a variant of the classical optimal transport (OT) distance [24], which measures the minimal cost of transporting mass between two distributions. Unlike the traditional Wasserstein distance [24], Sinkhorn distance introduces a regularization term, which makes the computation more efficient and helps avoid numerical instability issues. Mathematically, given two probability distributions μ and ν the Sinkhorn distance is defined as:

$$S_\epsilon(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)}\left(\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) + \epsilon H(\gamma)\right), \tag{5}$$

where $\Pi(\mu,\nu)$ is the set of joint distributions that satisfy the marginal constraints, and $c(x,y) = \|x-y\|^2$ is the cost function. $H(\gamma)$ is the entropy of the joint distribution γ, defined as:

$$H(\gamma) = -\int_{\mathcal{X}\times\mathcal{Y}}\gamma(x,y)\log(\gamma(x,y))\,dx\,dy, \tag{6}$$

The regularization parameter ϵ controls the strength of the entropy term, balancing the transportation cost and the entropy. Therefore, we adopt Sinkhorn distance to measure the discrepancy between $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_q$, which is represented as:

$$\mathcal{L}^{feat}_{DAS} = S_\epsilon(\boldsymbol{Z}_p,\boldsymbol{Z}_q), \tag{7}$$

Logits-level Distillation. To further recover the missing sentiment semantics, we constrain the logits of the two networks. The representations $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_q$ pass through fully connected layers and softmax layer to get logits $\boldsymbol{L}_p$ and $\boldsymbol{L}_q$. The Jensen-Shannon (JS) divergence is used as the measure of discrepancy, which serves as a symmetrical metric for assessing the similarity between two probability distributions, expressed as:

$$\mathcal{L}^{logits}_{DAS} = \mathcal{D}_{JS}(\boldsymbol{L}_p||\boldsymbol{L}_q)) = \frac{1}{2}(\mathcal{D}_{KL}(\boldsymbol{L}_p|\boldsymbol{M})) + \mathcal{D}_{KL}(\boldsymbol{L}_q||\boldsymbol{M})), \tag{8}$$

where $\boldsymbol{M}$ is the average distribution of $\boldsymbol{L}_p$ and $\boldsymbol{L}_q$. Ultimately, the loss of DAS is denoted as:

$$\mathcal{L}_{DAS} = \mathcal{L}^{feat}_{DAS} + \mathcal{L}^{logits}_{DAS}. \tag{9}$$

$\boldsymbol{L}_p$ and $\boldsymbol{L}_q$ are used to calculate the task loss $\mathcal{L}_{task}$. In regression and classification tasks. The task losses are Mean Squared Error (MSE) and cross-entropy losses, respectively. Eventually, the overall optimization objective $\mathcal{L}_{total}$ is expressed as $\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_{DHF} + \lambda_2\mathcal{L}_{DAS}$. where the hyperparameters λ_1 and λ_2 are 0.2 and 0.1.

III. Experiments

A. Datasets and Implementation Details

Datasets. The MOSI [25] dataset is a widely used benchmark for MSA, comprising 2,199 opinion video clips. It is divided into 1,284 clips for training, 229 for validation, and 686 for testing. MOSEI [26], an MSA dataset with 23,454 movie video clips, which is split into 16,326 training, 1,871 validation, and 4,659 testing samples. Each instance in both MOSI and MOSEI is assigned a sentiment label ranging from -3 (strongly negative) to +3 (strongly positive). For our evaluations, we compute the F1 score for positive/negative classification on these datasets.

The language modality is encoded into 768-dimensional vectors using the pre-trained BERT [16]. We utilize the COVAREP toolkit [27] to extract 74-dimensional acoustic features for the audio modality, which include 12 Mel-frequency cepstral coefficients, voiced/unvoiced segmentation features, and glottal source parameters. For the visual modality, we employ the Facet toolkit [28] to extract 35 facial action units that capture various facial movements.

Implementation details. Regarding the MOSI [25] and MOSEI [26] datasets, we use the aligned multimodal sequences therein as the original input for the FSRF. All models are implemented using the Pytorch [29] toolbox with four NVIDIA Tesla A800 GPUs. The Adam optimizer [30] is employed for network optimization. For the MOSI and MOSEI datasets, the hyper-parameters are set as follows: the learning rates are $\{1e-4, 2e-4\}$, the batch sizes are $\{16, 32\}$, the epoch numbers are $\{20, 30\}$. The embedding dimension is 128 on all two datasets. The missing features are replaced with zero values. To ensure fairness in our comparisons, we have reproduced several State-Of-The-Art (SOTA) methods and integrated them into our modality missing experimental setup. All experimental results are the average of the 10 random seed cases.

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

We select eight representative SOTA methods as baselines for comparison with FSRF. Specifically, to demonstrate the advantage of FSRF in the case of missing modality, we choose five missing-modality methods, including (1) joint learning methods, *i.e.*, MCTN [12], TransM [11], and CorrKD [15] and (2) generative methods *i.e.*, SMIL [31] and GCNet [9]. Additionally, to perform a more comprehensive comparison, we also select five complete-modality methods: Self-MM [5], CubeMLP [32], and DMD [7]. We analyze in detail the performance of the proposed FSRF and baseline models in the two modality-missing cases.

Analysis of inter-modality missingness. To simulate testing conditions of inter-modality missingness, we remove certain entire modalities from the samples. Tables I contrast the models' resilience to inter-modality missingness. We define "$\{l\}$" as the testing condition for both audio and visual modality are missing and "$\{l, a, v\}$" as the testing condition for the complete modality. "Avg. (Missing)" represents the mean performance of six testing conditions. We have the following key findings: (i) The inter-modality missingness leads to a decline in performance for all models, highlighting that combining complementary information from different modalities strengthens the sentiment semantics in joint representations. (ii) Under conditions where modalities are missing, the performance decline of methods using complete modalities is significantly greater than that of methods with missing modalities. This suggests that the missing-modality methods leverage their semantic-recovery learning approach to partially alleviate the impact of missing modalities. (iii) Among all the methods, our FSRF has the best performance, proving its strong robustness. For instance, on the MOSI dataset, FSRF's average F1 score is improved by 2.09% compared to CorrKD, and in particular by 3.92% in the testing condition $\{v\}$. The strength of this approach lies in its capacity to perform fine-grained representation factorization and execute precise bidirectional supervision. (iv) Different modalities have varying impacts on sentiment analysis. In unimodal situations, methods relying solely on the language modality achieve the best performance, with results similar to those obtained using all modalities. In bimodal cases, combinations involving the language modality outperform others, indicating that the language modality's rich, high-level semantics are crucial for recovering missing information.

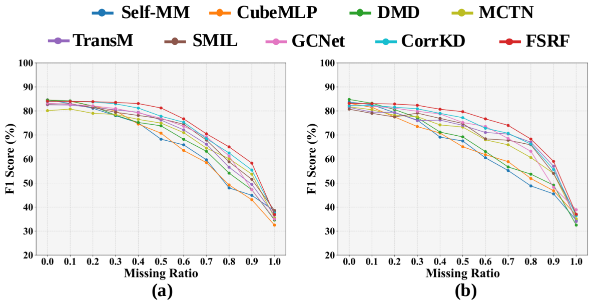

Fig. 2: Comparison results under intra-modality missingness cases on (a) MOSI and (b) MOSEI.

Analysis of intra-modality missingness. Intra-modality missingness is simulated by randomly discarding frame-level features from sequences at varying rates, with the missing ratio $p \in \{0.1, 0.2, \cdots, 1.0\}$. The performance curves of models with different p values are illustrated in Figure 2, offering an intuitive depiction of the robustness of all models. These results yield several key observations. (i) As the ratio p grows, every model's performance deteriorates. The finding suggests that intra-modality missingness induces considerable sentiment semantic loss and undermines the coherence of the integrated multimodal representations. (ii) In comparison to complete-modality methods, our FSRF demonstrates notable performance improvements under missing-modality conditions and achieves results that are competitive with those of complete modalities. (iii) Contrary to missing-modality methods, our FSRF demonstrates enhanced robustness. By factorizing modalities and applying high-dimensional distribution constraints within the self-distillation module, FSRF effectively recovers missing features and produces robust multimodal representations.

C. Ablation Studies

We validate the effects of proposed components in FSRF, including DHF and DAS. As illustrated in Table II, we conducted comprehensive ablation experiments of the two missing-modality cases in the MOSI and MOSEI datasets. The following observation proves the necessity of components: **(i)** First, we remove the DHF and concatenate the three original modalities in the DHF. The consistent performance degradation phenomenon in both missing modality cases shows that the proposed DHF can adequately capture the critical information and sentiment semantics in modalities. **(ii)** Then, when our DAS is eliminated, the worse performance demonstrates that performing bidirectional distributional alignment using

TABLE I: Comparison results under inter-modality missingness cases on MOSI and MOSEI.

Datasets	Models	Testing Conditions							
		$\{l\}$	$\{a\}$	$\{v\}$	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	Avg. (Missing)	$\{l,a,v\}$
MOSI	Self-MM [5]§	67.80	40.95	38.52	69.81	74.97	47.12	56.53	**84.64**
	CubeMLP [32]§	64.15	38.91	43.24	63.76	65.12	47.92	53.85	84.57
	DMD [7]§	68.97	43.33	42.26	70.51	68.45	50.47	57.33	84.50
	MCTN [12]‡	75.21	59.25	58.57	77.81	74.82	64.21	68.31	80.12
	TransM [11]‡	77.64	63.57	56.48	82.07	80.90	67.24	71.32	82.57
	SMIL [31]‡	78.26	67.69	59.67	79.82	79.15	71.24	72.64	82.85
	GCNet [9]‡	80.91	65.07	58.70	**84.73**	83.58	70.02	73.84	83.20
	CorrKD [15]‡	81.20	66.52	60.72	83.56	82.41	73.74	74.69	83.94
	FSRF (Ours)‡	**82.58***	**69.23***	**64.68***	84.19*	**83.72***	**76.25***	**76.78***	84.05*
MOSEI	Self-MM [5]§	71.53	43.57	37.61	75.91	74.62	49.52	58.79	83.69
	CubeMLP [32]§	67.52	39.54	32.58	71.69	70.06	48.54	54.99	83.17
	DMD [7]§	70.26	46.18	39.84	74.78	72.45	52.70	59.37	**84.78**
	MCTN [12]‡	75.50	62.72	59.46	76.64	77.13	64.84	69.38	81.75
	TransM [11]‡	77.98	63.68	58.67	80.46	78.61	62.24	70.27	81.48
	SMIL [31]‡	76.57	65.96	60.57	77.68	76.24	66.87	70.65	80.74
	GCNet [9]‡	80.52	66.54	61.83	81.96	81.15	69.21	73.54	82.35
	CorrKD [15]‡	80.76	66.09	62.30	81.74	81.28	71.92	74.02	82.16
	FSRF (Ours)‡	**82.62***	**69.12***	**65.04***	**82.27***	**81.35***	**74.98***	**75.90***	83.14*

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

TABLE II: Ablation results of FSRF on two datasets. "Intra-MM", "Inter-MM", and "CM": intra-modality missingness, inter-modality missingness, and complete modality cases.

(a) Ablation results on the MOSI dataset.

Models	Testing Conditions		
	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1 ↑
FSRF	**73.39**	**76.78**	**84.05**
w/o DHF	70.94	73.31	82.92
w/o DAS	71.36	74.83	83.11

(b) Ablation results on the MOSEI dataset.

Models	Testing Conditions		
	Intra-MM Avg. F1 ↑	Inter-MM Avg. F1 ↑	CM F1 ↑
FSRF	**72.94**	**75.90**	**83.14**
w/o DHF	70.43	73.81	82.51
w/o DAS	71.06	74.33	82.27

the self-distillation paradigm can generate favorable joint representations.

D. Qualitative Analysis

Figures 3(a) -3(c) show the visualization results of baseline models and Figure 3(d) presents the visualization results of our FSRF. The complete modality-based baseline model (*i.e.*, DMD [7]) cannot handle the problem of modality missingness, leading to extremely confusing latent distributions. The missing modality-based baseline models (*i.e.*, GCNet [9], and CorrKD [15]) mitigate the negative effects of modality missing to some extent and slightly distinguish the different categories. In contrast, the proposed method clearly separates the potential representations of samples with different sentiments and thus has the strongest performance and robustness.

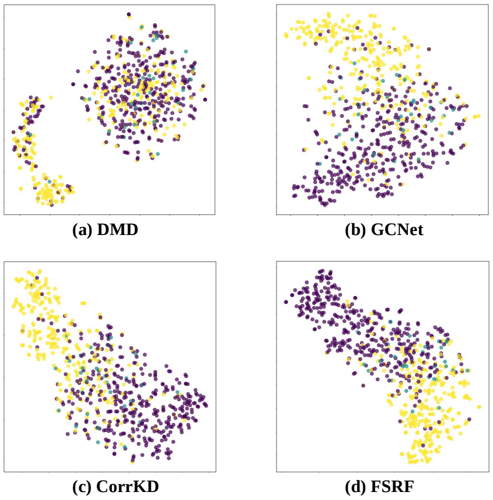
(a) DMD **(b) GCNet** **(c) CorrKD** **(d) FSRF**

Fig. 3: Representation visualization on the MOSI dataset.

IV. Conclusion

In this paper, we propose a Factorization-guided Semantic Recovery Framework (FSRF) to tackle various challenges posed by missing modalities in the MSA task. Specifically, we propose a de-redundant homo-heterogeneous factorization module that factorizes modality into modality-homogeneous, modality-heterogeneous, and noisy representations and design elaborate constraint paradigms for representation learning. Furthermore, we introduce a distribution-aligned self-distillation module that leverages bidirectional knowledge transfer under consistency constraint paradigms to effectively recover missing semantics. Comprehensive experiments validate the superiority of FSRF.

References

[1] Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang, "Robust emotion recognition in context debiasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024.

[2] Zhuyang Xie, Yan Yang, Jie Wang, Xiaorong Liu, and Xiaofan Li, "Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[3] Jiajia Tang, Dongjun Liu, Xuanyu Jin, Yong Peng, Qibin Zhao, Yu Ding, and Wanzeng Kong, "Bafn: Bi-direction attention based fusion network for multimodal sentiment analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1966–1978, 2022.

[4] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[5] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 10790–10797.

[6] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang, "Emotion recognition for multiple context awareness," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 144–162.

[7] Yong Li, Yuanzhi Wang, and Zhen Cui, "Decoupled multimodal distilling for emotion recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6631–6640.

[8] Dingkang Yang, Mingcheng Li, Linhao Qu, Kun Yang, Peng Zhai, Song Wang, and Lihua Zhang, "Asynchronous multimodal video sequence fusion via learning modality-exclusive and-agnostic representations," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[9] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao, "Gcnet: graph completion network for incomplete multimodal learning in conversation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[10] Yuanzhi Wang, Zhen Cui, and Yong Li, "Distribution-consistent modal recovering for incomplete multimodal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22025–22034.

[11] Zilong Wang, Zhaohong Wan, and Xiaojun Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, 2020, pp. 2514–2520.

[12] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 6892–6899.

[13] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," in *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 2022, pp. 1545–1554.

[14] Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang, "A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 10074–10082.

[15] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua 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.

[16] Jacob Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[18] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4400–4407.

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[20] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So 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.

[21] Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro 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.

[22] Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick, "On the importance of single directions for generalization," *Stat*, vol. 1050, pp. 15, 2018.

[23] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[24] Cédric Villani et al., *Optim. Transport Old New*, vol. 338, Springer, 2009.

[25] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[26] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe 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.

[27] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* IEEE, 2014, pp. 960–964.

[28] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "Openface: an open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[30] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng, "Smil: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 2302–2310.

[32] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin, "Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3722–3729.