Disentangled variational auto-encoder for multimodal fusion performance analysis in multimodal sentiment analysis

Accepted for publication in Knowledge-Based Systems.

Disentangled variational auto-encoder for multimodal fusion performance analysis in multimodal sentiment analysis

Rongfei Chen [a], Wenju Zhou [a], Huosheng Hu [b], Zixiang Fei [c], Minrui Fei [a], Hao Zhou [d]

[a] Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200444, China
[b] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK
[c] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[d] Department of Computer Science, University of Oxford, Oxford OX1 2JD, UK

Abstract

Multimodal Sentiment Analysis (MSA) holds extensive applicability owing to its capacity to analyze and interpret users' emotions, feelings, and perspectives by integrating complementary information from multiple modalities. However, inefficient and unbalanced cross-modal information fusion substantially undermines the accuracy and reliability of MSA models. Consequently, a critical challenge in the field now lies in effectively assessing the information integration capabilities of these models to ensure balanced and equitable processing of multimodal data. In this paper, a Disentanglement-based Variable Auto-Encoder (DVAE) is proposed for systematically assessing fusion performance and investigating the factors that facilitate multimodal fusion. Specifically, a dis-tribution constraint module is presented to decouple the fusion matrices and generate multiple low-dimensional and trustworthy disentangled latent vectors that adhere to the authentic unimodal input distribution. In addition, a combined loss term is modified to effectively balance inductive bias, signal reconstruction, and distribution constraint items to facilitate the optimization of neural network weights and parameters. Utilizing the proposed evaluation method, we can evaluate the fusion performance of multimodal models by contrasting the classifi-cation degradation ratio derived from disentangled hidden representations and joint representations. Experi-ments conducted with eight state-of-the-art multimodal fusion methods on the CMU-MOSEI and CMU-MOSEI benchmark datasets demonstrate that DVAE is capable of effectively evaluating the effects of multimodal fusion. Moreover, the comparative experimental results indicate that the equalizing effect among various advanced mechanisms in multimodal sentiment analysis, as well as the single-peak characteristic of the ground label distribution, both contribute significantly to multimodal data fusion.

Keywords: Multimodal sentiment analysis Model performance evaluation Disentangled representation learning

1. Introduction

Multimodal sentiment analysis (MSA) is a crucial area in affective computing. By leveraging cross-modal information and feature integration, MSA models bridge the gap between vision, speech, and language. As shown in Fig. 1-①, these models improve the accuracy of predicting sentiments by integrating various signals. Consequently, MSA models are widely used in healthcare [1,2], intelligent education [3], and social opinion monitoring [4]. Although numerous state-of-the-art fusion strategies in MSA have achieved excellent experimental results, there remain challenges in assessing the effectiveness and credibility of multimodal fusion strategies. The modal bias in multimodal fusion (shown in Fig. 1-②) impedes cross-modal semantic complementarity between different modal representations, hindering models from generating reliable joint representations across different datasets [5–7]. Moreover, inappropriate coupling in cross-modal scenarios amplifies the instability of end-to-end model training, directly increasing the uncertainty in the performance evaluation of multimodal fusion models [8]. Therefore, exploring a robust approach for collectively evaluating models, particularly addressing modality bias and inappropriate coupling in multimodal fusion, has garnered significant attention among researchers in recent years.

Currently, performance metrics like accuracy, precision, recall, and F1 score are commonly used to evaluate multimodal fusion [9]. However, relying solely on these metrics is insufficient due to the limitation of assessing the correlation between unimodal vectors and multimodal joint matrices [10]. Additionally, techniques like Layer-wise Relevance Propagation (LRP) [11,12] have been used to visualize the contributions of different modal representations. The effectiveness of these methods is still affected by the architecture and configuration of the neutral network [11]. This sensitivity becomes more apparent when integrating multiple feature extraction components and fusion units with different mechanisms into a multimodal fusion model. Some state-of-the-art techniques, such as attention-based assessment methods [13] and parameter optimization methods [14], are dedicated to exploring inter-modal information interactions and feature dependencies by optimizing weight assignments [15].

However, their interpretability and operating efficiency are greatly limited by inherent black-box nature or additional gradient computation [6], thus reducing the generalization ability of evaluation methods. Following the groundbreaking advancements in Disentangled Representation Learning (DRL) that transforms high-dimensional, entangled features into low-dimensional explanatory elements [16], a great number of DRL-based approaches are widely employed to evaluate the fusion performance in multimodal learning [17,18]. Nevertheless, current DRL-based techniques focus on refining generated features through information bottlenecks without considering the distribution gap arising from multimodal fusion, making it more likely to produce modally irrelevant representations and thereby potentially reducing the validity and reliability of multimodal information fusion.

In this context, there is a pressing demand for an efficient and trustworthy methodology to evaluate multimodal fusion performance in MSA. In this paper, a model evaluation approach centered on disentangled representation, named Disentangled Variational Auto-encoder (DVAE), is proposed in Multimodal Sentiment Analysis (MSA). DVAE simulates the multimodal fusion process by decoupling and reconstructing joint representation, aiming to evaluate and investigate the determinants affecting multimodal fusion by analyzing discrepancies in the distribution of different representations (shown in Fig. 1-③). Notably, the distribution constraint layers guide encoders in generating disentangled latent unimodal representations. This layer helps to bridge the distribution gap between disentangled latent representations derived from joint representations and unimodal representations employed for multimodal fusion. Additionally, a modified combined loss term is constructed to incorporate distribution constraint loss, reconstruction loss, and inductive bias loss, aiming to balance the parameter optimization of the neural network. The main contributions of this paper can be summarized as follows:

- A novel DVAE model is proposed to evaluate the effectiveness of multimodal fusion in MSA by introducing decoupled representation learning.
- A modality distribution constraint layer is proposed to guide the encoder in decoupling joint representations and generating explainable disentangled latent representations. In addition, a modified combined loss term is presented to balance and facilitate the parameters update during model training.
- Experimental results characterizing distributional properties indicate that the equilibrium effect between recurrent neural units and attentional mechanisms, as well as the unimodal nature of ground label distributions, can enhance multimodal information fusion.

The rest of the paper is organized as follows. Section II outlines some related work in terms of multimodal fusion strategies, explainable

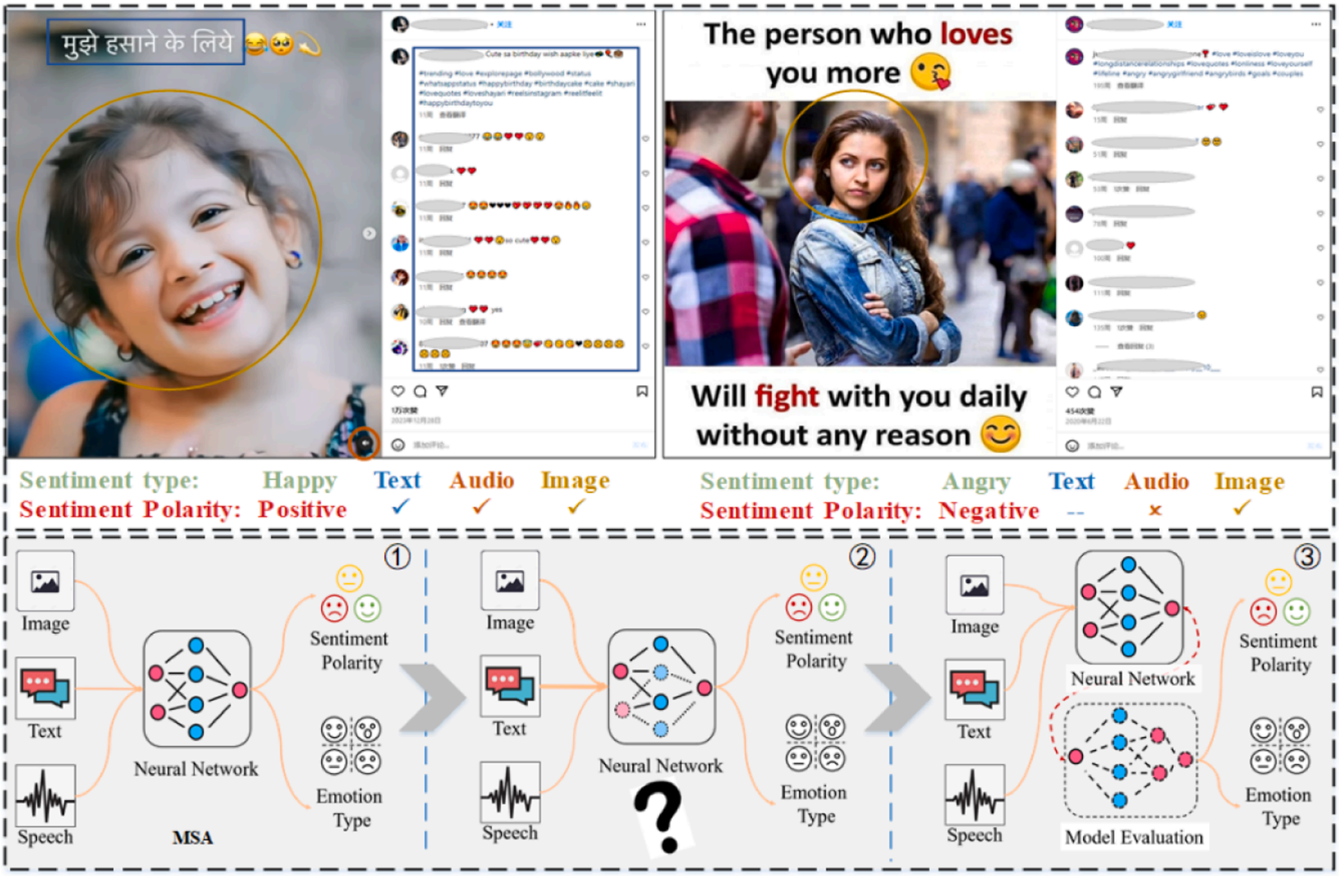


Fig. 1. Schematic analysis of MSA. The upper section of the figure shows an example of the Instagram (Ins) scene related to multimodal sentiment analysis. The lower section includes the schematic diagram of the fusion model (①), the schematic diagram illustrating the difficulties in model evaluation (②), and the proposed model evaluation method depicted in the schematic diagram (③).

evaluation techniques, and disentangled representation learning. In Section III, our proposed methodology is presented, including foundation and notation, an overview of DVAE, and a principle analysis of disentanglement. Section IV describes the experiment setup such as datasets, evaluation metrics, baselines, and evaluation setup in experiments. Then, experimental results and analysis are given in Section V, including experiments for evaluating disentangled representations and experiments for evaluating multimodal fusion. Finally, a brief conclusion and future work are presented in Section VI.

2. Related work

2.1. Multimodal information fusion

Multimodal information fusion techniques can enhance model performance by capturing consistent correlations among heterogeneous multi-granularity features [19]. Consequently, an efficient multimodal fusion strategy is essential for multimodal sentiment analysis tasks [20]. Typically, multimodal fusion techniques are categorized into three types, namely feature-level fusion (early fusion), decision-level fusion (late fusion), and model-level fusion (hybrid fusion) [21].

The feature-level fusion approaches focus on directly feeding the extracted multimodal discriminating features into classifiers for category prediction. For example, Amir Zadeh et al. [22] proposed the multi-attention recurrent network for human communication comprehension, aiming to discover potential interactions between modalities by leveraging discriminated modal dynamics and novel neural components. However, concatenating features from different modalities in feature-level fusion models typically results in high-dimensional joint representations, thereby increasing computational complexity and the risk of overfitting [23]. Additionally, the feature-level fusion approach struggles with the asynchronous nature of different modal data, leading to inaccurate predictor results. In contrast, the decision-level fusion approaches achieve superior classification results by reweighting the decision vectors of unimodal discriminators. Nevertheless, these models are constrained by inadequate cross-modal information interaction, diminishing the complementarity of multimodal affective features [23]. Model-level fusion methods address these limitations by integrating intermediate representations generated by different encoders to improve classification accuracy. However, these methods also increase model complexity and make the training process more challenging [24].

In addition, advanced mechanisms, including attention-based techniques [25,26] and recurrent cell-based neural networks [27], have been employed in multimodal sentiment analysis tasks to investigate dependencies among cross-modal features. For example, Xingye Li et al. [28] proposed an Expectation-maximized Cross-modal Temporal (ECT) fusion approach to capture interactions and long-term dependencies in visual, audio, and textual data. Changqin Huang et.al [29] presented a Text-centered Fusion Network with cross-modal Attention (TeFNA) to effectively models unaligned multimodal timing information by incorporating cross-modal attention with mutual information.

2.2. Multimodal fusion performance evaluation

Numerous fusion performance assessment techniques in MSA, encompassing metric comparisons and data visualization, are undergoing continuous evolution [30]. These tools provide valuable insights into multimodal fusion in MSA, guiding model construction and feature selection in multimodal sentiment analysis [31]. For example, Sandeep et al. [32] performed a detailed empirical evaluation of modern approaches, including LSTM, RNN, CNN, and CapsNet, for semantic analysis, and then assessed the model performance by utilizing standard classification metrics such as precision, recall, accuracy, AUC, and F1 score. Recently, there has been a gradual increase in the adoption of multimodal evaluation strategies employing explainable AI techniques. These strategies focus on quantifying model performance and qualitatively analyzing multimodal fusion process [33,34]. However, metrics- and visualization-based evaluation methods are inadequate for addressing model bias and fairness in multimodal information fusion, thus making it difficult to provide valid and interpretable performance evaluation results.

A number of studies have evaluated the performance of multimodal models by examining the functions and contributions of key components within these models [35]. For instance, Qinghua Zhao et al. [36] investigate a distinct approach to handling input items and their weights by designing a neural structure. This structure not only learns a discriminative representation of the target task via its encoder but also concurrently monitors key elements through its localizer. Given the compatibility between different components and the integrity of the model, the aforementioned evaluation methods potentially faced with challenges in assessing the impact of global feature changes on the overall model performance. Therefore, multimodal feature engineering, which aims to explore the relationship between feature quality and model performance, has been developing rapidly. For example, a series of comparative experiments were conducted in [37] to investigate the contributions of several common word embeddings to sentiment classification models. Ao Feng et al. [38] conduct comprehensive experiments on various network components, including different word embeddings and convolutional kernels, to highlight the significance of these components in evaluating the model performance.

2.3. Disentangled representation learning in MSA

Disentangled representation learning is an unsupervised technique that separates each feature into narrowly defined variables with distinct dimensions [39]. In Multimodal Sentiment Analysis, this method captures key affective information from complex representations and generates explainable outputs for handling challenging real-world tasks [40,41]. For instance, Dingkang Yang et al. [40] proposed a feature-separated multimodal recognition method that learns common and private features for each modality by mapping input data to modality-invariant and modality-specific subspaces. Similarly, Imant Daunhawer et al. [42] introduced a novel multimodal generative model designed to capture the joint distribution across multiple modalities. This model combines modality-specific and shared factors, efficiently aggregating shared information from any subset of modalities.

Currently, disentangled representation learning is increasingly employed in Multimodal Sentiment Analysis to investigate interpretable affective representations [43,44]. For example, Dr. Emotion [45], an integrated framework implemented by separating and disentangling the implicitly encoded emotions from the content in latent space, is constructed to learn disentangled representations of social media posts (i.e., tweets) for emotion analysis. In Multimodal Sentiment Analysis (MSA), while disentangled representation learning can decouple the fusion matrix to obtain independent modal components, these components may struggle to accurately capture the statistical distributional information of input data. Additionally, evaluation metrics for individual modalities may not be able to directly quantify multimodal fusion performance, potentially leading to deviations between evaluation criteria and actual model performance. Therefore, research on multimodal sentiment analysis based on disentangled representation learning has increasingly focused on reducing statistical errors in data decoupling, intending to enhance the generalization ability of decoupled models. Yuhao Zhang et al. [46] put forward a disentangled sentiment representation adversarial network to mitigate domain shifts of expressive styles in multimodal cross-domain sentiment analysis, aiming to improve the adaptability of multimodal models.

3. Methodology

3.1. Formulation and notion

In this section, we present the Variable Auto-Encoder (VAE) as a framework to briefly introduce the core concept of disentangled representation learning. Assume the latent space variable $\mathscr{z}$ is a vector following a Gaussian distribution $\mathscr{z} \sim \mathscr{N}(0, I)$, which is employed for generating the observed $\mathscr{X} = \{x_1,\ x_2,\ x_i,\ \cdots,\ x_n\}$. The primary goal of the VAE-based disentangled representation learning method is to learn the parametric encoder $p(\mathscr{z}|x)$ for maximizing the log-likelihood of observations (i.e., $p(x) = \mathbb{E}_{p(z)}[p(x|z)]$) with introducing the prior $p(\mathscr{z})$ and the likelihood $p(x|z)$ of generating x given z. Due to the computational complexity of calculating $p(x)$, resulting in an intractable distribution $p(\mathscr{z}|x)$. Thus, the variational inference was introduced to address the intractable problem. In other word, an approximate posterior distribution $q_\phi(\mathscr{z}|x)$ implemented by networks with parameters ϕ is employed as an approximation of the intractable true posterior $p(\mathscr{z}|x)$. Moreover, the sum of the maximized log-likelihood $\mathscr{L}_{max} = \sum log[p(x)]$ serves as the optimization objective in VAE, encouraging the network to achieve the unbiased reconstruction of inputs. Formally, the log-likelihood function can be written as following formulas with the approximate posterior $q_\phi(\mathscr{z}|x)$.

$$\mathscr{L}_{max} = \mathscr{L}_{ELBO} + \mathscr{D}_{KL}\left(q_\phi(\mathscr{z}|x)\ \|\ p(\mathscr{z}|x)\right) \tag{1}$$

where $\mathscr{L}_{ELBO}$ is empirically called the Evidence Lower Bound (*ELBO*), which is formally defined as the sum of $\mathbb{E}_{q_\phi(z|x)}[log(p(x|z)p(z))]$ and $\mathscr{H}(z)$. Specifically, $\mathscr{H}(z) = -\mathbb{E}_{q_\phi(z|x)}\left[\log\left(q_\phi(z|x)\right)\right]$. The $\mathscr{D}_{KL}\left(q_\phi(\mathscr{z}|x)\ \|\ p(\mathscr{z}|x)\right)$ is the relative entropy used to measure the difference in distribution between $q_\phi(\mathscr{z}|x)$ and $p(\mathscr{z}|x)$. Due to the non-negative property of $\mathscr{D}_{KL}$, the inequality $\mathscr{L}_{max} \geq \mathscr{L}_{ELBO}$ consistently holds. In addition, the $\mathscr{L}_{ELBO}$ can be rewritten as $\mathbb{E}_{q_\phi(z|x)}\left[log\left[p(z)/q_\phi(z|x)\right] + log[p(x|z)]\right]$, according to the conditional probability. Based on the conversion illustrated in Eq. (2), the maximum likelihood estimated parameters can be translated into the output of the neural network.

$$\begin{cases} \mathbb{E}_{q_\phi(z|x)}\left[\log\dfrac{q_\phi(z|x)}{p(z)}\right]\ \sum\limits_{i}^{n}\left(e^{\sigma_i} - (1+\sigma_i) + \mu_i^2\right) \\ \qquad\qquad \mathbb{E}_{q_\phi(z|x)}[log[p(x|z)]]\ \mathscr{L}_{nn} \end{cases} \tag{2}$$

where $\mathscr{L}_{nn}$ is the loss function of a neural network. μ and σ are the mathematical expectations and variances of the approximate posterior, respectively. Numerous VAE-based approaches have been proposed for implementing disentanglement by incorporating supervised meta-priors. Therefore, disentanglement can be implemented by utilizing simple networks, which opens the opportunity to disentangle explainable low-dimensional latent representations from coupled joint representations fused by multimodal fusion approaches.

3.2. Overview of disentangled variational auto-encoder

The Disentangled Variational Auto-encoder can achieve disentanglement of joint representations by generating independent explainable latent representations that capture the statistical distribution characteristics of a single modality. By generating distinct representations for each modality from the joint representation, the approach enhances the clarity and comprehensibility of the fused representation in multimodal sentiment analysis, preserving the original properties and patterns of each data type. In particular, DVAE can evaluate the efficacy of multimodal fusion by comparing performance variations among disentangled latent vectors, original joint representations, and reconstructed joint representations. It evaluates the effectiveness of the fusion model in fusing multimodal information by quantifying the distinctions among the unimodal features, joint representations, and reconstructed joint representations. For example, it can assess the extent to which unimodal representations of a special modality encapsulate identical information as the fused representation by calculating the mutual information indicators, as well as the degree of alignment between the reconstructed representation and the original data through comparing classification metrics.

The framework structure of the proposed method is shown in Fig. 2, which consists of a fusion module and a disentanglement module. The fusion module has three parts: the video inputs with visual frames I_v, speech utterances I_a, and text subtitles I_t, the extracted unimodal features $\mathscr{M}_a$, $\mathscr{M}_v$ and $\mathscr{M}_t$, and a pre-trained multimodal fusion model $\mathscr{F}_{a,v,t}$ with coupled joint representations $\mathscr{X}_\mathscr{J}$. It is important to highlight that our objective is to assess the fusion impact of multimodal models in MSA and investigate the crucial factors influencing this effect. Therefore, the model structure, hyperparameter settings, and data processing of the fusion module all utilize existing pre-trained models and parameter configurations. For the disentanglement module, a disentanglement-based unsupervised learning network is designed as a fundamental framework to map joint representations into explainable independent modality-related vectors with separate dimensions. Specifically, the three independent encoders F_a, F_v and F_t equipped with distributional constraints $\mathscr{C}_\Omega^{a/t/v}$ are proposed to generate independent modality-related latent representations from a coupled joint representation $\mathscr{X}_\mathscr{J}$. Moreover, the weight-shared decoder $\widetilde{F}_{a,v,t}$ focus on outputting the reconstructed representation $\mathscr{X}'_\mathscr{J}$ based on the concatenated latent vectors $(\mathscr{Z}_a, \mathscr{Z}_v, \mathscr{Z}_t)$. The disentanglement module serves three primary purposes: 1. It indirectly evaluates the efficacy of fusing multimodal pre-trained models by contrasting the performance decay of joint representations and hidden-variable representations on the same classifier. 2. To evaluate the fusion performance of pre-trained fusion models, the level of multimodal fusion is quantified by calculating the degradation ratios of the multimodal fusion model with the simple concatenation method in the MSA classification metrics 3. Based on the disentanglement module, we seek to identify the key factors influencing multimodal fusion by examining the distribution of features and weight assignments within this module.

According to the network structure, the objective function of DVAE is composed of $\mathscr{L}_\mathscr{M}^{a|v|t}$, $\mathscr{L}_\mathscr{Z}^{a|v|t}$, and $\mathscr{L}_{rec}^{a,v,t}$. A detailed description and discussion of the mathematical principle of this part will be elaborated later. We assume that the statistical distribution of any isolated unimodal feature $\mathscr{M}_a$, $\mathscr{M}_v$ or $\mathscr{M}_t$ complies with the independent identically distributed property ($\mathscr{M}_* \sim \mathscr{N}(\mu_*, \sigma_*)$). During network training, the DVAE module maps the coupled matrices (i.e. joint representations) into the independent modality-related vectors, i.e. $F_a(\mathscr{X}_\mathscr{J}) \rightarrow \mathscr{Z}_a$, $F_v(\mathscr{X}_\mathscr{J}) \rightarrow \mathscr{Z}_v$, and $F_t(\mathscr{X}_\mathscr{J}) \rightarrow \mathscr{Z}_t$. Different from the standard VAE that fed reconstruction data $\mathscr{X}_\mathscr{J}'$ to classifiers, the disentangled latent representations with independent modality-related vectors $\mathscr{Z}_{a,v,t} = concat(\mathscr{Z}_a,\ \mathscr{Z}_v,\ \mathscr{Z}_t)$ and the reconstructed joint representation $\mathscr{X}_\mathscr{J}'$ are fed into the classifier $\Phi(*)$ for evaluating the multimodal fusion performance.

3.3. Principle analysis of disentangled variational auto-encoder

Disentangled representation learning, recognized as an interpretable learning tool, has found widespread application in machine learning tasks such as image generation and has demonstrated notable success in enhancing feature interpretability and reducing data dimensionality. However, the general AE-based frameworks for achieving disentanglement are limited in the context of multimodal fusion. This limitation arises from the disparities in distribution that exist between different modalities within a multimodal domain, which differs from the traditional application of disentangled representation learning in image-based interpretable analysis. Thus, we incorporate distributional

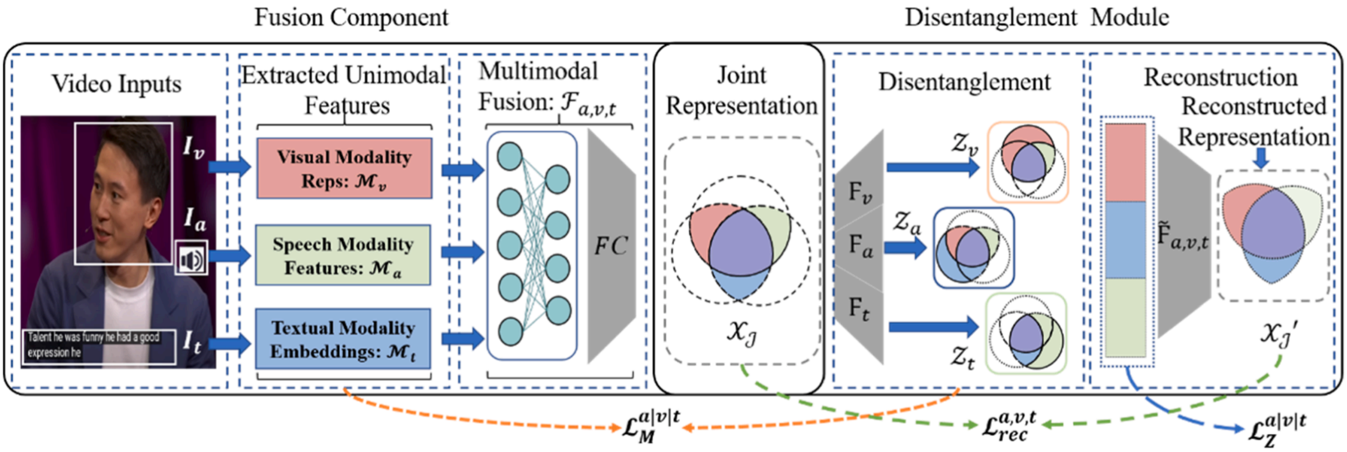

Fusion Component
Disentanglement Module
Video Inputs
I_v
I_a
I_t
Talent he was funny he had a good expression he
Extracted Unimodal Features
Visual Modality Reps: $\mathcal{M}_v$
Speech Modality Features: $\mathcal{M}_a$
Textual Modality Embeddings: $\mathcal{M}_t$
Multimodal Fusion: $\mathcal{F}_{a,v,t}$
FC
Joint Representation
$\mathcal{X}_{\mathcal{J}}$
Disentanglement
F_v
F_a
F_t
Z_v
Z_a
Z_t
Reconstruction
Reconstructed Representation
$\tilde{\mathrm{F}}_{a,v,t}$
$\mathcal{X}_{\mathcal{J}}'$
$\mathcal{L}_M^{a|v|t}$
$\mathcal{L}_{rec}^{a,v,t}$
$\mathcal{L}_Z^{a|v|t}$

Fig. 2. The model architecture of DVAE and performance evaluation framework.

constraint terms $\mathscr{C}_{\Omega}^{a/t/v}$ to direct each independent encoder in producing unimodal latent variables $\mathscr{Z}_a/\mathscr{Z}_v/\mathscr{Z}_t$ that align with the target distribution. To specify the disentanglement module of DVAE, the disentanglement and reconstruction of the joint representation are shown in Fig. 3.

As shown in Fig. 3., the joint representations $\mathscr{X}_{\mathscr{J}}$ are firstly fed into the independent encoders $F_a/F_v/F_t$. Then, each encoder generates the unimodal latent representations $\mathscr{Z}_a/\mathscr{Z}_v/\mathscr{Z}_t$ consistent with the distribution of original input single modalities $\mathscr{M}_a/\mathscr{M}_v/\mathscr{M}_t$. Notably, we assume each variable $\mathscr{M}_a/\mathscr{M}_v/\mathscr{M}_t$ follows the Gaussian distribution $\mathscr{M}_* \sim \mathscr{N}(\mu_*,\ \sigma_*)$, which gives the reason that $\mathscr{Z}_a \sim \mathscr{N}(\mu_a,\ \sigma_a)/\mathscr{Z}_v \sim \mathscr{N}(\mu_v,\ \sigma_v)/\mathscr{Z}_t \sim \mathscr{N}(\mu_t,\ \sigma_t)$ shown in Fig. 2. Next, the reconstructed joint representation $\mathscr{X}'_{\mathscr{J}}$ is obtained from the concatenated latent representation $\mathscr{Z}_{a,v,t}$ using the shared decoder $\widetilde{F}_{a,v,t}$. Based on the process of decoupling and reconstructing the data, we illustrate the efficacy of different modules by comparing the metrics discrepancy across different representations, and ultimately demonstrate the rationality and effectiveness of the method in fusion performance evaluation: 1. Illustrating the effectiveness of the DRL in disentangling and reconstructing fused data by comparing the classification metrics between $\mathscr{X}_{\mathscr{J}}$ and $\mathscr{X}'_{\mathscr{J}}$ on the MSA task. 2. Demonstrating the capability of the modality constraint layer in generating specific distributional features through an experimental results comparison of $\mathscr{Z}_a/\mathscr{Z}_v/\mathscr{Z}_t$ with $\mathscr{M}_a/\mathscr{M}_v/\mathscr{M}_t$. 3. Evaluating the fusion performance of a model by comparing the degradation ratio (i.e., $|(Metric_{updated} - Metric_{baseline})\ /Metric_{baseline}| * 100\%$ %) between $\mathscr{Z}_{a,v,t}$ and $\mathscr{X}_{\mathscr{J}}$

In addition, we present a detailed mathematical analysis of DVAE from the perspective of objective optimization. As shown in Fig. 4, the optimization objective of a standard VAE includes the inductive bias $\mathscr{L}_{\mathscr{Z}}^{a|v|t}$ (i.e. $\mathscr{D}_{KL}\left(q_{\phi}(z|x)\ \|\ p(z)\right)$ and signal reconstruction item $\mathscr{L}_{rec}^{a,v,t}$(i.e. $\mathbb{E}_{q_{\phi}(z|x)}[log[p(x|z)]]$), Considering the distribution discrepancy of each modality during multimodal feature fusion, the intermediate distribution constraint term layer is embedded into standard VAE to maintain consistent statistical distribution between $(\mathscr{M}_a,\ \mathscr{M}_v,\ \mathscr{M}_t)$ and $(\mathscr{Z}_a,\ \mathscr{Z}_v,\ \mathscr{Z}_t)$. Thus, the optimization objective of DVAE is updated as an adaptive combination of $\mathscr{L}_{\mathscr{Z}}^{a|v|t}$, $\mathscr{L}_{rec}^{a,v,t}$and an augmented constraint item $\mathscr{L}_{\mathscr{M}}^{a|v|t}$. Different from the encoding flow in standard VAE that directly extracts latent factors from original inputs, DVAE generates explainable modality-related disentangled latent representations consistent with the distribution of isolated unimodal features by employing distribution constraint $\mathscr{C}_{\Omega}^{a/t/v}$, bridging the information distribution gap in multimodal fusion analysis.

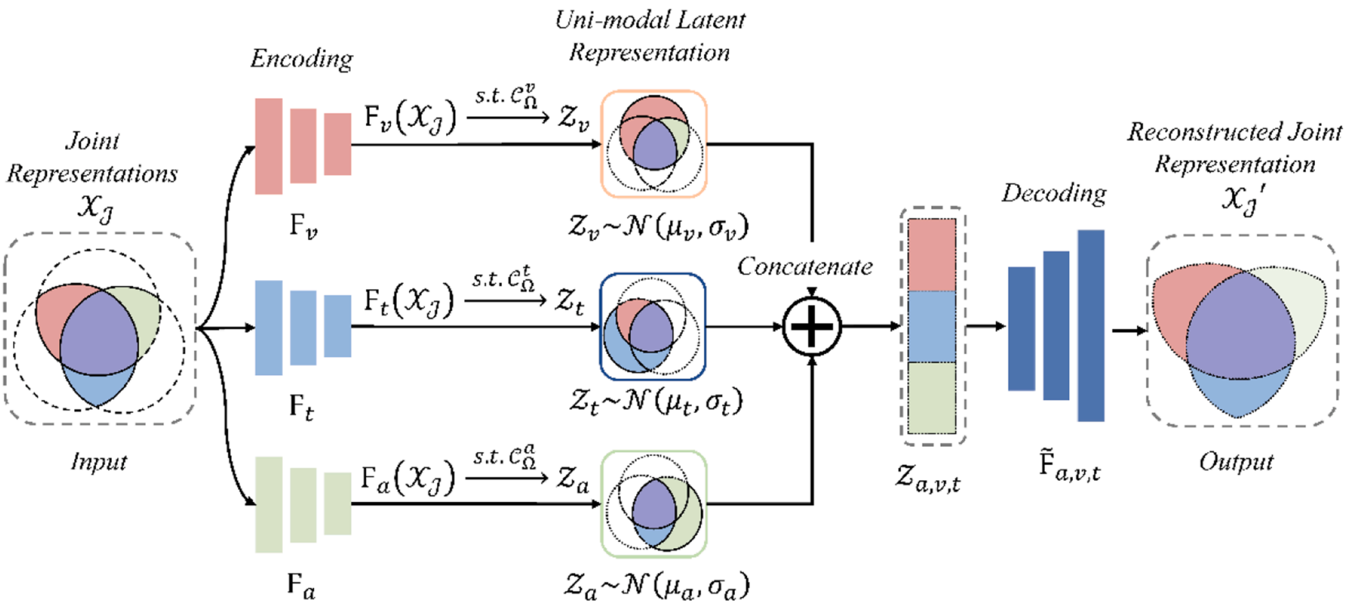

Encoding
Uni-modal Latent Representation
Joint Representations $\mathcal{X}_{\mathcal{J}}$
Input
F_v
F_t
F_a
$\mathrm{F}_v(\mathcal{X}_{\mathcal{J}}) \xrightarrow{s.t.\ \mathcal{C}_{\Omega}^{v}} Z_v$
$\mathrm{F}_t(\mathcal{X}_{\mathcal{J}}) \xrightarrow{s.t.\ \mathcal{C}_{\Omega}^{t}} Z_t$
$\mathrm{F}_a(\mathcal{X}_{\mathcal{J}}) \xrightarrow{s.t.\ \mathcal{C}_{\Omega}^{a}} Z_a$
$Z_v \sim N(\mu_v, \sigma_v)$
$Z_t \sim N(\mu_t, \sigma_t)$
$Z_a \sim N(\mu_a, \sigma_a)$
Concatenate
$Z_{a,v,t}$
Decoding
$\tilde{\mathrm{F}}_{a,v,t}$
Reconstructed Joint Representation $\mathcal{X}_{\mathcal{J}}'$
Output

Fig. 3. The disentanglement and reconstruction of the joint representation in the disentanglement module.

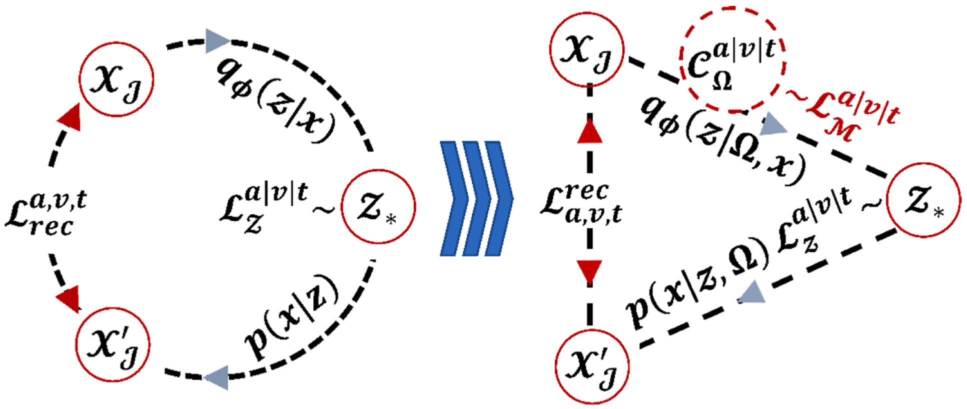

$\mathcal{X}_\mathcal{J}$
$q_\phi(z/x)$
$\mathcal{L}_{rec}^{a,v,t}$
$\mathcal{L}_Z^{a|v|t} \sim$
$\mathcal{Z}_*$
$\mathcal{X}'_\mathcal{J}$
$p(x|z)$
$\mathcal{X}_\mathcal{J}$
$\mathcal{C}_\Omega^{a|v|t}$
$\sim \mathcal{L}_\mathcal{M}^{a|v|t}$
$q_\phi(z|\Omega, x)$
$\mathcal{L}_{a,v,t}^{rec}$
$\mathcal{Z}_*$
$p(x|z,\Omega)\ \mathcal{L}_Z^{a|v|t} \sim$
$\mathcal{X}'_\mathcal{J}$

Fig. 4. Mathematical analysis of the Disentangled Variational Auto-encoder.

As embedding the distribution constraint $\mathscr{C}_\Omega^{a/t/v}$, the approximate posterior $q_\phi(\mathscr{z}|x)$ and the likelihood $p(x|\mathscr{z})$ of standard VAE are upgraded to $q_\phi(\mathscr{z}|\Omega, x)$ and $p(x|\mathscr{z},\ \Omega)$, where Ω denotes the intermediate modality variables. Therefore, the log-likelihood function presented in Eq. (1) can be extended as:

$$\mathscr{L}_{max} = \sum log[p(x)] = \frac{1}{2}\cdot\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}[log[p(x)]] + \frac{1}{2}\cdot\mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}[\log[p(x)]] \qquad (3)$$

where the first term can be formulated as:

$$\frac{1}{2}\cdot\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}[log[p(x)]]$$

$$= \frac{1}{2}\cdot\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}\left[log\left[\frac{p(x,\ \Omega,\mathscr{z})}{p(\mathscr{z},\ \Omega|x)}\cdot\frac{q_\phi(\mathscr{z}|\Omega, x)}{q_\phi(\mathscr{z}|\Omega, x)}\right]\right]$$

$$= \frac{1}{2}\cdot\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}\left[log\left[\frac{q_\phi(\mathscr{z}|\Omega, x)}{p(\mathscr{z},\ \Omega|x)}\right] + log\left[\frac{p(x,\ \Omega,\mathscr{z})}{q_\phi(\mathscr{z}|\Omega, x)}\right]\right] \qquad (4)$$

Similarly,

$$\frac{1}{2}\cdot\mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}[log[p(x)]]$$

$$= \frac{1}{2}\cdot\mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}\left[\log\left[\frac{p(x,\ \Omega,\mathscr{z})}{p(\mathscr{z},\ \Omega|x)}\cdot\frac{q_\phi(\Omega|x,\mathscr{z})}{q_\phi(\Omega|x,\mathscr{z})}\right]\right]$$

$$= \frac{1}{2}\cdot\mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}[log[\frac{q_\phi(\Omega|x,\mathscr{z})}{p(\mathscr{z},\ \Omega|x)}] + \log[\frac{p(x,\ \Omega,\mathscr{z})}{q_\phi(\Omega|x,\mathscr{z})}]] \qquad (5)$$

Since,

$$\begin{cases} \mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)} log\left[\frac{q_\phi(\mathscr{z}|\Omega,\ x)}{p(\mathscr{z},\ \Omega|x)}\right] = \mathscr{D}_{KL}\left[q_\phi(\mathscr{z}|\Omega,\ x)\|\ p(\mathscr{z},\ \Omega|x)\right] \\ \mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})} log\left[\frac{q_\phi(\Omega|x,\mathscr{z})}{p(\mathscr{z},\ \Omega|x)}\right] = \mathscr{D}_{KL}\left[q_\phi(\Omega|x,\mathscr{z})\|\ p(\mathscr{z},\ \Omega|x)\right] \end{cases} \qquad (6)$$

are non-negative constants. Therefore, the maximum log-likelihood estimation considering the distribution constraint can be expressed as:

$$\mathscr{L}_{max} \geq \mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}\left[log\left(\frac{p(\mathscr{z},\ \Omega, x)}{q_\phi(\Omega|x,\mathscr{z})}\right)\right] + \mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}\left[log\left(\frac{p(\mathscr{z},\ \Omega, x)}{q_\phi(\mathscr{z}|\Omega,\ x)}\right)\right] \qquad (7)$$

where the polynomial on the right-hand side of the inequality denotes the improved Evidence Lower Bound (ELBO).

The fundamental concept behind distribution constraint relies on an inductive bias: Assume that the distributions of the isolated unimodal features($\mathscr{M}_a,\ \mathscr{M}_v,\ \mathscr{M}_t$) and disentangled latent variables ($\mathscr{Z}_a,\ \mathscr{Z}_v,\ \mathscr{Z}_t$) follow the Gaussian distribution $p(\mathscr{M}_a) \sim \mathscr{N}(\mu_a,\ \ \sigma_a)\ \ /\ p(\mathscr{M}_v) \sim \mathscr{N}(\mu_v,\ \ \sigma_v)/p(\mathscr{M}_t) \sim \mathscr{N}(\mu_t,\ \ \sigma_t)$and $p(\mathscr{Z}_a) \sim \mathscr{N}(0,\ \ 1)/p(\mathscr{Z}_v) \sim \mathscr{N}(0,\ \ 1)/p(\mathscr{Z}_t) \sim \mathscr{N}(0,\ \ 1)$, respectively. Under this assumption, we can convert the intractable a priori optimization problem into a solvable parametric optimization by training neural networks to learn all probability distribution variables in the *ELBO*.

Based on the principles, we can proceed with further optimization for variable Ω:

$$\mathbb{E}_{q_\phi(\Omega|x,z)}\left[log\left(\frac{p(z,\ \Omega, x)}{q_\phi(\Omega|x, z)}\right)\right] = \underbrace{\mathbb{E}_{q_\phi(\Omega|x,z)} log\left(p(\Omega)/q_\phi(\Omega|x, z)\right)}_{(1)}$$
$$+\underbrace{\mathbb{E}_{q_\phi(\Omega|x,z)} log(p(z|\Omega))}_{(2)} + \underbrace{\mathbb{E}_{q_\phi(\Omega|x,z)} log(p(x, z|\Omega))}_{(3)} \qquad (8)$$

where ① is $-\mathscr{D}_{KL}\left[q_\phi(\Omega|x,\mathscr{z})\ \ \|\ \ p(\Omega)\right]$, and ② as well as ③ measure the effects of embedding modal distribution constraint layers on generating modality-related variables ($\Omega\rightarrow\mathscr{Z}$) and signal reconstruction ($\Omega\rightarrow\mathscr{Z}\rightarrow\mathscr{X}'_\mathscr{J}$). Since the signal reconstruction includes the latent variable generation (i.e. $\{\Omega\rightarrow\mathscr{Z}\} \in \{\Omega\rightarrow\mathscr{Z}\rightarrow\mathscr{X}'_\mathscr{J}\}$), Eq. (8) can be simplified as:

$$\mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}\left[\log\left(\frac{p(\mathscr{z},\ \Omega, x)}{q_\phi(\Omega|x,\mathscr{z})}\right)\right] = \mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})}\log\left(\frac{p(\Omega)}{q_\phi(\Omega|x,\mathscr{z})}\right)$$
$$+ \mathbb{E}_{q_\phi(\Omega|x,\mathscr{z})} log(p(x,\mathscr{z}|\Omega)) \qquad (9)$$

Similarly, for latent variable $\mathscr{Z}$:

$$\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}\left[log\left(\frac{p(\mathscr{z},\ \Omega, x)}{q_\phi(\mathscr{z}|\Omega,\ x)}\right)\right] = \mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}\log\left(\frac{p(\mathscr{z})}{q_\phi(\mathscr{z}|\Omega,\ x)}\right)$$
$$+ \mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)} log(p(\Omega|\mathscr{z}))$$
$$+ \mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)} log(p(x|\mathscr{z},\ \Omega)) \qquad (10)$$

Specifically, $\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)}\log\left(p(\mathscr{z})/q_\phi(\mathscr{z}|\Omega,\ x)\right)$ encourages each modality-related encoder to output statistical parameters that are consistent with the distribution of isolated unimodal features. $\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)} log(p(x|\mathscr{z},\ \Omega))$ is used for evaluating the performance of data reconstruction of each modality ($\Omega\rightarrow\mathscr{Z}\rightarrow\mathscr{X}'_\mathscr{J}$). Notably, the flow of generation from Ω to $\mathscr{Z}$ is an irreversible process because the reconstruction is directly decoded by the latent representations without considering the reconstruction process $\mathscr{Z}\rightarrow\Omega$, i.e., $\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)} log(p(\Omega|\mathscr{z})) = log[p(\Omega)] = C$.

Thus, Eq. (10) can be simplified as the sum of the two remaining items $-\mathscr{D}_{KL}\left[q_\phi(\mathscr{z}|\Omega,\ x)\ \ \|\ \ p(\mathscr{z})\right]$ and $\mathbb{E}_{q_\phi(\mathscr{z}|\Omega,\ x)} log(p(x|\mathscr{z},\ \Omega))$, which are consistent with that of standard VAE. The complete *ELBO* of maximizing the log-likelihood of observed $\mathscr{X}_\mathscr{J}$ is denoted as:

$$\begin{aligned}\mathscr{L}_{ELBO} = & -\mathscr{D}_{KL}\left[q_{\phi}(\mathscr{z}|\Omega,\ x)\|\ p(\mathscr{z})\right] + \mathbb{E}_{q_{\phi}(\mathscr{z}|\Omega,\ x)}log(p(x|\mathscr{z},\ \Omega)) \\ & -\mathscr{D}_{KL}\left[q_{\phi}(\Omega|x,\mathscr{z})\|\ p(\Omega)\right] + \mathbb{E}_{q_{\phi}(\Omega|x,\mathscr{z})}log(p(x|\mathscr{z},\ \Omega))\end{aligned} \tag{11}$$

Since $\mathscr{L}_{ELBO}$ determines the lower bound of $\mathscr{L}_{max}$, the total objective optimization $\mathscr{L}_{max}$ can be degraded to parameter estimation of the $\mathscr{L}_{ELBO}$. Furthermore, it is noteworthy that there exists a connection between maximum likelihood estimation and neural network optimization, as depicted in Eq. (2). Thus, we can theoretically determine the optimal estimation parameters with the remarkable nonlinear fitting capability of neural networks. The optimization items can be converted to the combination of loss functions following the mapped rule illustrated in Eq.12 and Eq.13,

$$\begin{cases}\mathscr{D}_{KL}\left[q_{\phi}(\mathscr{z}|\Omega,\ x)\|\ p(\mathscr{z})\right] \rightleftharpoons \dfrac{1}{2}\left(log\sigma^{2} - \left(\mu^{2}+\sigma^{2}\right) + 1\right) \\ s.t.\ q_{\phi}(\mathscr{z}|\Omega,\ x) \sim \mathscr{N}(\mu_{1},\sigma_{1}),\ p(\mathscr{z}) \sim \mathscr{N}(0,1)\end{cases} \tag{12}$$

$$\begin{cases}\mathscr{D}_{KL}\left[q_{\phi}(\Omega|x,\mathscr{z})\|\ p(\Omega)\right] \rightleftharpoons \dfrac{1}{2}\left(log\dfrac{\sigma_{1}^{2}}{\sigma_{2}^{2}} - \dfrac{\sigma_{1}^{2}}{\sigma_{2}^{2}} - \left(\dfrac{\mu_{1}^{2}-\mu_{2}^{2}}{\sigma_{2}^{2}}\right) + 1\right) \\ s.t.\ q_{\phi}(\Omega|x,\mathscr{z}) \sim \mathscr{N}(\mu_{2},\sigma_{2}),\ p(\mathscr{z}) \sim \mathscr{N}(\mu_{1},\sigma_{1})\end{cases} \tag{13}$$

The top formula is the inductive bias of latent variables, which aims to measure the similarity between the estimated distribution $q_{\phi}(\mathscr{z}|\Omega,\ x)$ using neural networks and the true distribution of latent variables $p(\mathscr{z})$. Likewise, the bottom one is used to calculate the similarity of $q_{\phi}(\Omega|x,\mathscr{z})$ and $p(\Omega)$. Notably, the intermediate variable $p(\Omega)$ is an unbiased estimate obtained by computing the statistics of each isolating input modality while $q_{\phi}(\Omega|x,\mathscr{z})$ is an estimated value fitted by neural networks. The remaining terms $\mathbb{E}_{q_{\phi}(\mathscr{z}|\Omega,\ x)}log(p(x|\mathscr{z},\ \Omega))$ and $\mathbb{E}_{q_{\phi}(\Omega|x,\mathscr{z})}log(p(x|\mathscr{z},\ \Omega))$, which represents the comprehensive encoding ($\mathscr{X}_{\mathscr{J}}\rightarrow\Omega\rightarrow\mathscr{Z}$) and decoding ($\Omega\rightarrow\mathscr{Z}\rightarrow\mathscr{X}'_{\mathscr{J}}$) of the complete process ($\mathscr{X}_{\mathscr{J}}\rightarrow\mathscr{X}'_{\mathscr{J}}$) in DVAE, can be formally defined as $\mathscr{L}_{rec}^{a,v,t}$, following the definition of standard VAE. Therefore, the objective function $\mathscr{L}_{total}$ of DVAE can be viewed as an optimization of the standard VAE function $\mathscr{L}_{VAE}$ under the constraint of the function $\mathscr{L}_{\mathscr{M}}^{a|v|t}$,

$$\mathscr{L}_{total} = \mathscr{L}_{\mathscr{M}}^{a|v|t} + \mathscr{L}_{VAE} \tag{14}$$

where $\mathscr{L}_{\mathscr{M}}^{a|v|t}$ is the distribution constraint loss for assessing the distribution similarity, which can be implemented by calculating the *KL divergence*. $\mathscr{L}_{VAE}$ is the objective function of standard VAE, which includes $\mathscr{L}_{\mathscr{Z}}^{a|v|t}$ and $\mathscr{L}_{rec}^{a,v,t}$.

4. Experimental studies

To investigate the viability of assessing model performance through the proposed method and to explore the fundamental factors that govern the efficacy of multimodal fusion, we empirically conducted several comparative experiments on MSA by utilizing the popular benchmark datasets, namely CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [47], CMU Multimodal Corpus of Sentiment Intensity (CMU-MOSI) [48], The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [49] and The CHinese SIngle- and Multimodal Sentiment analysis dataset (CH-SIMS) [50]. Furthermore, several advanced techniques, including memory modules, attention mechanisms, and recurrent cell-based components, serve as comparative approaches to demonstrate the exceptional capability of DVAE in generating disentangled and explainable representations, as well as in evaluating the effects of multimodal fusion.

4.1. Datasets

The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset comprises 2199 opinion video clips, each accompanied by sentiment annotations within the $[-3, 3]$ range. This dataset undergoes meticulous annotation, encompassing subjectivity, sentiment intensity, as well as per-frame and per-opinion annotated visual features, along with audio features annotated at the per-millisecond level. The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset is the largest dataset of multimodal sentiment analysis and emotion recognition to date. This gender-balanced dataset comprises over 23,500 sentence utterance videos randomly selected from various topics and monologue videos, featuring more than 1000 online YouTube speakers. During the experiments, all the utterances are divided into training (16265 samples), validating (1869 samples), and test sets (4693 samples) where each utterance is labeled with a ratio score from -3 (highly negative) to 3 (highly positive). The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) comprises 151 recorded dialogue videos, totaling 302 videos. Each clip is labeled with nine emotions (anger, excitement, fear, sadness, surprise, frustration, happiness, disappointment, and neutral), as well as potency, arousal, and dominance. The CHinese SIngle- and Multimodal Sentiment analysis dataset (CH-SIMS) contains 60 videos with 2281 discourses gathered from movies, TV series, and variety shows. The average length of each discourse is 3.67 s, and each video retains only the speaker's facial image. Each discourse is assigned one multimodal label and three unimodal labels per video. The labels in this dataset are: positive, weak positive, neutral, weak negative, and negative.

4.2. Metrics

Motivated by the evaluation work in [51], several evaluation metrics for illustrating the feasibility of disentanglement and the effectiveness of multimodal fusion are chosen on benchmark datasets: *Binary accuracy*: $\mathscr{A}cc_{2}$, which determines the sentiment polarity based on positive and negative values. Moreover, we adopted the *Multiple Classification Accuracy:* $\mathscr{A}cc_{7}$ to evaluate the sentiment intensity within a range of [-3,3], with values of -3 and 3 representing extreme negatives and positives respectively. Furthermore, the *F1 score*: $\mathscr{F}1$ and the *Mean Absolute Error*: $\mathscr{M}ae$ are used for error analysis while *Pearson's correlation*: $\mathscr{R}$ measures the information relationship between predictions and ground truth labels. Specifically, we adopted the $\|\ \mathscr{x}\|_{1}$ to measure the distribution discrepancy between original unimodal inputs and disentangled latent modality-related vectors in all experiments. In addition, we measure the fusion performance of the multimodal model by calculating the performance degradation ratios.

4.3. Baselines

Based on the sophisticated fundamental structures such as attention mechanism and contextual awareness function, these advanced models are categorized broadly into the following categories:

Attention-based methods (Abm): The utilization of attention mechanism components for fusing multimodal representations is prominently observed in the following advanced attention-based techniques: Multimodal Transformer network *Multi-Transformer (MulT)* [52], *Multi--attention Recurrent Network (MARN)* [22], and *Recurrent Attended Variation Embedding Network (RAVEN)* [53]. Through several stacked attention components, these models combine the most relevant multimodal dynamics to provide a joint representation for category prediction in MSA.

Recurrent unit-based models (Rubm): By utilizing stacked recurrent unit-based components, *Recurrent Neural Networks (RNNs)* can effectively aggregate multimodal sentiment information throughout the specified time series. Based on the proposed performance evaluating strategy, this study investigates various multimodal fusion models based on the recurrent unit, namely *Early-Fusion LSTM (EF-LSTM)* [54], *Late-fusion LSTM (LF-LSTM)* [54], *Recurrent Multistage Fusion Network (RMFN)* [55], as well as *Long-Short Term Hybrid Memory (LSTHM)* [22], to evaluate feasibility of disentanglement and assess the fusion effectiveness of the recurrent unit-based approaches in MSA.

In addition, several innovative DNN-based models have become essential components in multimodal fusion, which rely on the characterization of special data structures, such as *Memory-based Fusion (MFN)* [56], *Late Fusion using DNN (LF-DNN)* [57], *Low-rank Multimodal Fusion (LMF)* [58], and *Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA)* [59]. Specifically, MFN is a memory fusion network consisting of several multimodal gated memory components, which can output multimodal sentence representations by building loops between the memory units, hidden state units, and output units of the LSTMs. LF-DNN is a sequence learning method that utilizes input-level feature fusion and bi-directional long short-term memory (BLSTM) deep neural networks (DNNs). It is designed to identify the sentiment type and intensity of multimodal data at the input level; Low-rank multimodal fusion (LMF) is a method to make multimodal fusion efficient without compromising performance using a low-rank power tensor; MISA is a multimodal sentiment analysis model utilizing domain adaptive learning, which projects each modality into both modality-invariant and modality-specific subspaces. This approach aims to learn common information shared across multimodal data while capturing the unique features of each modality.

4.4. Evaluation setup

Our architecture consists of three modules: multimodal fusion models, the disentanglement module, and the simple classifier (e.g., the MLP). Following the operation flow from multimodal fusion models to the classifier, we aim to explore the feasibility of multimodal disentanglement in the case of optimal performance of each module. Firstly, we trained the various multimodal fusion approaches and froze the optimal architecture configuration and network parameters, which ensured that each multimodal fusion model could generate high-quality joint representations with relatively advanced performance in MSA. Secondly, a network embedding of representation disentanglement is implemented by concatenating DVAE to the output layer of the pre-trained model, aiming to generate the independent modality-related vectors and disentangled latent representations. Similarly, we linked a simple but efficient network MLP as the common classifier after freezing the pre-trained combined network (multimodal fusion models with embedded disentanglement modules) to predict the sentiment values in the range of [−3, 3].

Additionally, the *Early Stopping Strategy* is employed to prevent model over-fitting and the *Stochastic Gradient Descent* technique (e.g. the *Adam optimizer*) is used to update model parameters. Notably, we freeze the optimal model parameters of pre-trained multimodal fusion networks before disentanglement and reconstruction to reduce the risk of parameter interference that may exist during multiple model training. To avoid over-fitting the network training, the default training epoch is 100 with early stop patience setting *patience=20*. For the benchmark data, all data parameters and environment configurations are consistent with that in [51]. We trained all models in the PyTorch framework with a learning rate of 0.001, batch size of 256, and 100 epochs on a hardware cluster equipped with an NVIDIA GeForce GTX 1660Ti GPU and an Intel (R) Core (TM) i7–9750H CPU, achieving the best performance on the benchmark datasets. The training, validation, and testing sets are derived from the same sampling and segmentation strategy to ensure comparability and conformity. We consistently use a straightforward single-layer perception model (MLP) as the unique classifier. The output structure of the fusion model was finely fine-tuned to additionally output the latent representations produced by the fusion process, which is subsequently fed into DVAE. Since the loss term in DVAE encompasses distinct sub-loss terms with disparate functions, the *Automatic Weighted Loss* technique proposed in [57] is employed for optimizing losses at each epoch.

5. Results and discussion

Within this section, our attention is directed toward two experiments and the subsequent analysis of experimental findings—the experiment on evaluating disentangled representations and assessing multimodal fusion. Notably, the evaluating experiment of disentanglement aims to verify that the performance of the joint representation and its reconstructed outcome in the MSA task remains essentially identical within acceptable error limits, providing a basis for further evaluation and analysis of the fusion performance.

5.1. Experiments for evaluating disentangled representations

Our primary objective is to demonstrate the feasibility of using disentangled representation learning for multimodal model performance evaluation. In this section, we experimentally demonstrate the feasibility of decoupled representation learning from two perspectives: 1. Classification Comparison Experiments. We demonstrate the effectiveness of the proposed disentangled variational auto-encoder for disentangling and reconstructing joint representations by comparing the performance similarity between joint and decoupled representations on sentiment classification tasks; 2. Convergence Analysis Experiments. We demonstrate the stability of the decoupled variational auto-encoder model by visualizing the loss function trend. Notably, all parameters and running environments of the multimodal fusion model are consistent with the settings in [51].

Classification Comparison Experiments: We quantified the disparity between the reconstructed representations and the joint representations on CMU-MOSEI and CMU-MOSI benchmark datasets. The comparative results on the CMU-MOSEI dataset are presented in Table 1. We've organized the baselines into distinct categories based on the model's core mechanism for result comparison. It's important to mention that " *Rubm** " denotes a variant of the *Recurrent unit-based models (Rubm)*. We can see that RMFN exhibits superior performance in $\mathscr{A}cc_2$: 78.61 % / 72.51 %, $\mathscr{A}cc_7$: 46.16 % / 48.48 %, $\mathscr{F}_1$: 77.10 % / 59.55 %, and $\mathscr{R}$: 0.600 for both joint representations and reconstruction representations. For $\mathscr{M}ae$, RAVEN reports the optimal value of 0.666 in terms of joint representations, as compared to 0.646 for RMFN. Then, the classification results on the CMU-MOSI dataset are illustrated in Table 2. The RAVEN demonstrates superior classification performance on the CMU-MOSI benchmark dataset, with accuracy scores of 76.64 % / 73.97 % for $\mathscr{A}cc_2$, 32.22 % / 31.63 % for $\mathscr{A}cc_7$, the lowest value of 1.014/1.020 for $\mathscr{M}ae$, the strongest correlation of 0.617 for $\mathscr{R}$, and a precision-recall trade-off value of 74.39 % / 73.73 %. The experimental results reveal that the classification performance of the joint representation is very similar to that of the reconstructed

Table 1
Classification results on the CMU-MOSEI dataset (**Joint Representations / Reconstructed Representations**).

Method	Type	$\mathscr{A}cc_2$(%)	$\mathscr{A}cc_7$(%)	$\mathscr{F}$1(%)	$\mathscr{M}ae$	$\mathscr{R}$
EF-LSTM	Rubm	69.07/ 71.03	45.34/ 41.37	70.29/ 59.00	0.696/ 0.841	0.542/ 0.005
LF-LSTM	Rubm	66.42/ 70.98	44.76/ 41.37	67.91/ 59.02	0.774/ 0.838	0.527/ **0.060**
RMFN	Rubm	**78.61**/ **72.51**	46.16/ **48.48**	**77.10**/ **59.55**	0.670/ **0.646**	**0.600**/ 0.040
LSTHM	Rubm	64.81/ 71.05	45.55/ 41.35	66.44/ 59.00	0.712/ 0.837	0.528/ 0.040
MulT	Abm	71.06/ 71.03	42.21/ 41.37	70.85/ 59.00	0.768/ 0.841	0.384/ 0.005
MARN	Abm	77.80/ 71.03	46.00/ 41.37	76.28/ 59.0	0.677/ 0.838	0.576/ 0.044
RAVEN	Abm	77.13/ 71.03	47.83/ 41.37	76.54/ 59.00	**0.666**/ 0.842	0.590/ 0.005
MFN	Rubm*	77.23/ 71.03	**48.30**/ 41.37	76.83/ 59.00	0.674/ 0.838	0.576/ 0.008

Table 2
Classification results on the CMU-MOSI dataset (**Original Representations / Reconstructed Representations**).

Method	Type	$\mathscr{A}cc_2$(%)	$\mathscr{A}cc_7$(%)	$\mathscr{F}1$(%)	$\mathscr{M}ae$	$\mathscr{R}$
EF-LSTM	Rubm	73.03/ 72.59	30.76/ 30.17	73.10/ 72.67	1.049/ 1.069	0.592/ 0.030
LF-LSTM	Rubm	72.89/ 72.74	30.03/ 28.57	72.93/ 72.79	1.052/ 1.062	0.584/ 0.049
RMFN	Rubm	73.18/ 73.76	30.90/ 29.88	73.07/ 73.62	1.022/ 1.034	0.604/ 0.017
LSTHM	Rubm	70.41/ 70.99	26.38/ 27.55	70.36/ 70.96	1.129/ 1.150	0.534/ **0.058**
MulT	Abm	60.50/ 59.62	24.49/ 24.34	59.40/ 58.37	1.352/ 1.339	0.343/ 0.034
MARN	Abm	72.01/ 71.87	31.63/ 33.24	71.90/ 71.78	1.074/ 1.080	0.574/ 0.031
RAVEN	Abm	**74.64/ 73.97**	**32.22/ 31.63**	**74.39/ 73.73**	**1.014/ 1.020**	**0.617**/ 0.023
MFN	Rubm*	72.30/ 72.01	30.90/ 31.20	72.39/ 72.00	1.042/ 1.057	0.591/ 0.013

representation, particularly in terms of accuracy. This similarity suggests that the proposed model can generate reconstructed representations with performance comparable to joint representations, indirectly demonstrating the effectiveness of DVAE in disentangling and reconstructing joint representations. It is also worth noting that the differences between the two representations in terms of correlation metrics $\mathscr{R}$ are quite pronounced, as we will continue to analyze in *Section: Experiments for Evaluating Multimodal Fusion*.

Convergence Analysis Experiments: We conducted random comparative experiments on the benchmark datasets and visualized the network loss changes during model training and validation. The visualization results are shown in Fig. 4 and Fig. 5, respectively. Fig. 6

The experimental findings shown in Tables 1 and 2 demonstrate a slight degradation in the classification performance of the mentioned models on MSA when embedding DVAE into the multimodal fusion approach. A plausible explanation is that the extra distribution constraint item $\mathscr{L}_{\mathscr{M}}^{a|v|t}$ results in a higher convergence lower bound. Embedding DVAE into the multimodal fusion models results in a modification of the loss item from $\mathscr{L}_{vae}^{a|v|t}$ (i.e., $\mathscr{L}_{VAE}$, the sum of the latent loss and the reconstruction loss) to $\mathscr{L}_{\mathscr{M}}^{a|v|t} + \mathscr{L}_{vae}^{a|v|t}$ (i.e., $\mathscr{L}_{total}$, the total loss), where $\mathscr{L}_{\mathscr{M}}^{a|v|t}$ (i.e., the modality loss) is a positive semi-definite variable.

We can observe that the model-based loss curves of validation processes gradually stabilize and reach a minimum value. The small difference in magnitude between $\mathscr{L}_{total}$ and $\mathscr{L}_{vae}^{a|v|t}$ is likely attributed to the introduced loss item generated by the modality constraint layer. Moreover, the trend and magnitude of the modal loss term exhibited considerable variation across different datasets but remained consistent within each specific dataset. The likelihood of this phenomenon is influenced by both the dataset size and the richness of the provided features. It is noteworthy that in the CMU-MOSI dataset, the modal loss increases until it reaches a convergence point and then stabilizes. As the latent loss and the reconstruction loss exert a more significant influence on the overall loss during network training than the modal loss, the automatic weight assignment technique allocates substantial weights to data with high uncertainty and may lead to under-training of the modal loss function in the pre-training stage. Despite this, the convergence of the neural network total loss function is not significantly hindered by the modal loss function. Therefore, DVAE proves the efficacy of employing disentangled representation learning to decouple multimodal joint representations and generate reconstructed joint representations in multimodal fusion in multimodal sentiment analysis tasks, as evidenced by both the comparison of classification results and the analysis of loss function convergence.

5.2. *Experiments for evaluating multimodal fusion*

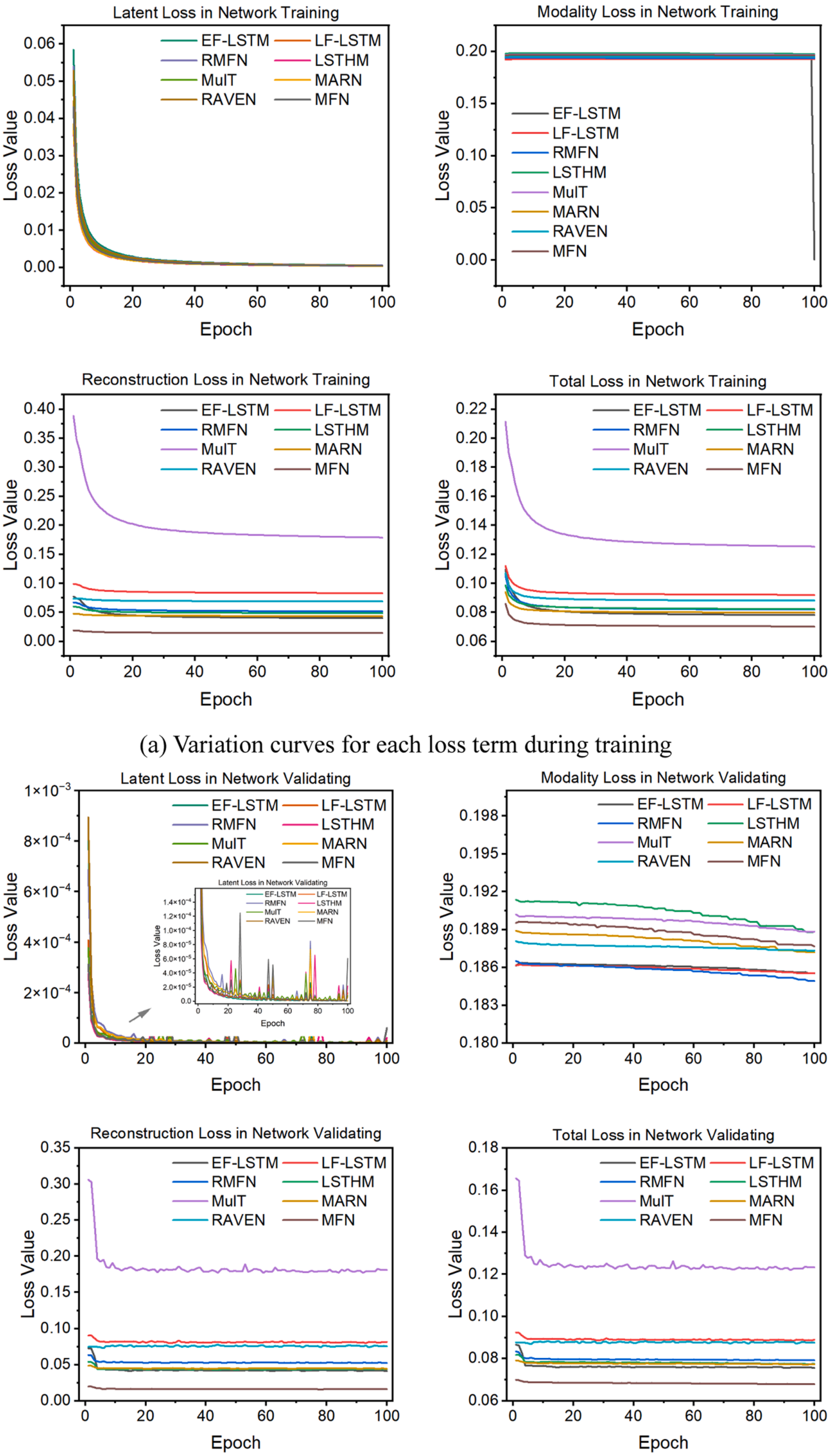
To quantify the effectiveness of different fusion strategies in fusing multimodal data, we measure the degradation ratios of classification metrics on benchmark datasets between the concatenated fusion and other state-of-the-art fusion methods. Then, we identify the dominant factors influencing the fusion effectiveness by visualizing the correlation between latent and joint representations. The fundamental concept of the DVE model involves substituting the original unimodal inputs with disentangled unimodal representations, as well as replacing the original joint representation with the disentangled joint representation. The objective is to assess the fusion impact of multimodal models by comparing the classification disparities between the original representations and disentangled representations in MSA. Thus, before reassessing the fusion experiments, we conducted two ablation experiments to eliminate potential interference from other variables on the experimental outcomes.

Original unimodal representation VS. Disentangled unimodal representations: To demonstrate the consistency between the unimodal latent representations produced by the modal constraint layer and the original input unimodal representations, we individually predicted the classification results for various unimodal inputs. Fig. 7 displays the result distributions of different unimodal representations from various fusion methods across diverse classification metrics. The figure reveals a subtle distinction in the distribution of classification outcomes between the decoupled unimodal representation and the original unimodal representation.

The prediction outcomes relying on unimodal representations exhibit superior classification accuracy in the CMU-MOSEI dataset compared to the CMU-MOSI dataset, which can be intuitively derived from the color comparisons of the accuracy metrics in the two datasets. Additionally, the MAE errors observed in the CMU-MOSEI dataset are lower than those observed in the CMU-MOSI dataset. The experimental results on the CMU-MOSEI dataset indicate that the classification performance of partially disentangled unimodal representations outperforms that of the original unimodal representations, with this improvement being especially notable in the textual modalities. Detailed discussions on the causes of this phenomenon will be provided in *Section: Experiments for Evaluating Multimodal Fusion*. For the CMU-MOSI dataset, it is evident that the classification performance of the disentangled unimodal representations closely matches that of the original unimodal representations in the MSA task. This observation demonstrates the effectiveness of the modal constraint layer in guiding the network to generate disentangled unimodal representations that align well with the distribution of the original unimodal representations. Based on the results analysis above, it can be demonstrated that disentangled unimodal representations can substitute the original unimodal inputs in prediction without compromising classification accuracy.

Unimodal latent representations VS. Concatenated latent representations: Besides the proposed classifier-based metrics for indirectly evaluating the quality of disentangled representations, the *Mutual Information Gap Score* (MIG Score) [60] has been suggested for evaluating the degree of disentanglement in representations during network training, thereby establishing the utility of disentangled representations to improve performance on downstream tasks.

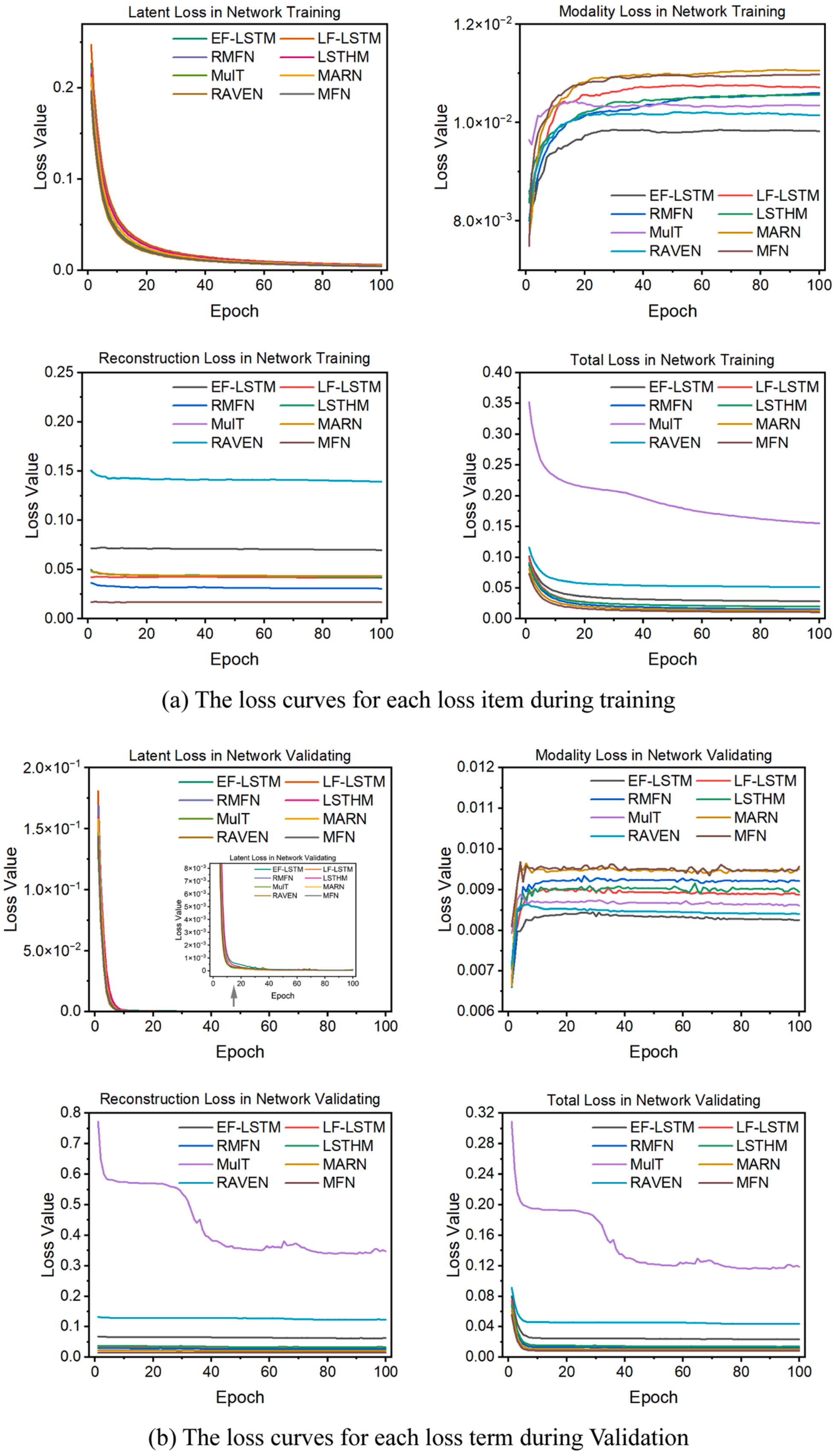
Following the work using the MIG score to isolate the independent factors of variation [61], we quantified the information correlation between the explainable disentangled latent vectors and coupled joint matrices using the fine-turned MIG Score. Instead of calculating the Mutual Information of ground truth factors in the implemented MIG score [60], the fine-turned MIG Score utilizes information entropy as a proxy of Mutual Information used in the original MIG Score, which is more consistent with the calculation of entropy of ground truth factors mentioned in MIG Score and more suitable for entropy calculation of multimodal heterogeneous data. As described in [61], the empirical mutual information between a latent variable and a ground truth factor



(a) Variation curves for each loss term during training

(b) Variation curves for each loss term during Validation

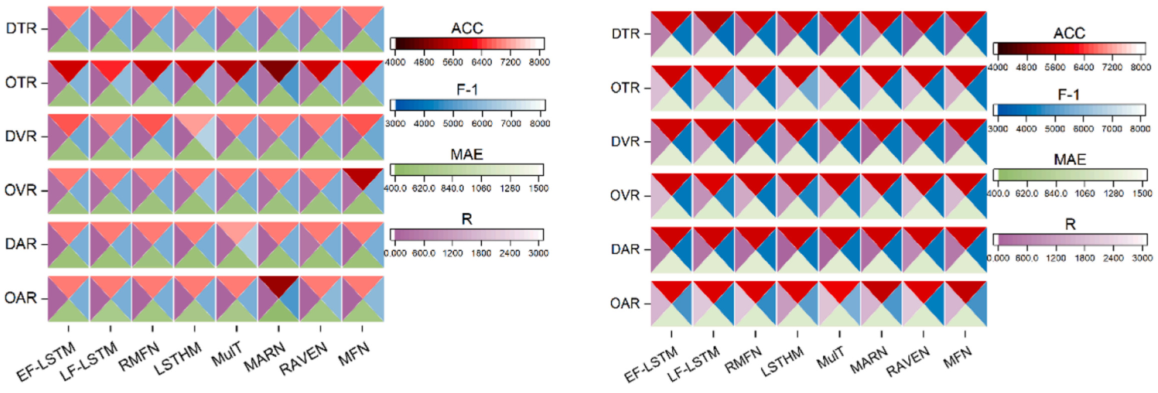
Fig. 5. Variation curves for each loss term of different models in the CMU-MOSEI dataset.



(a) The loss curves for each loss item during training

(b) The loss curves for each loss term during Validation

Fig. 6. Variation curves for each loss term of different models in the CMU-MOSI dataset.



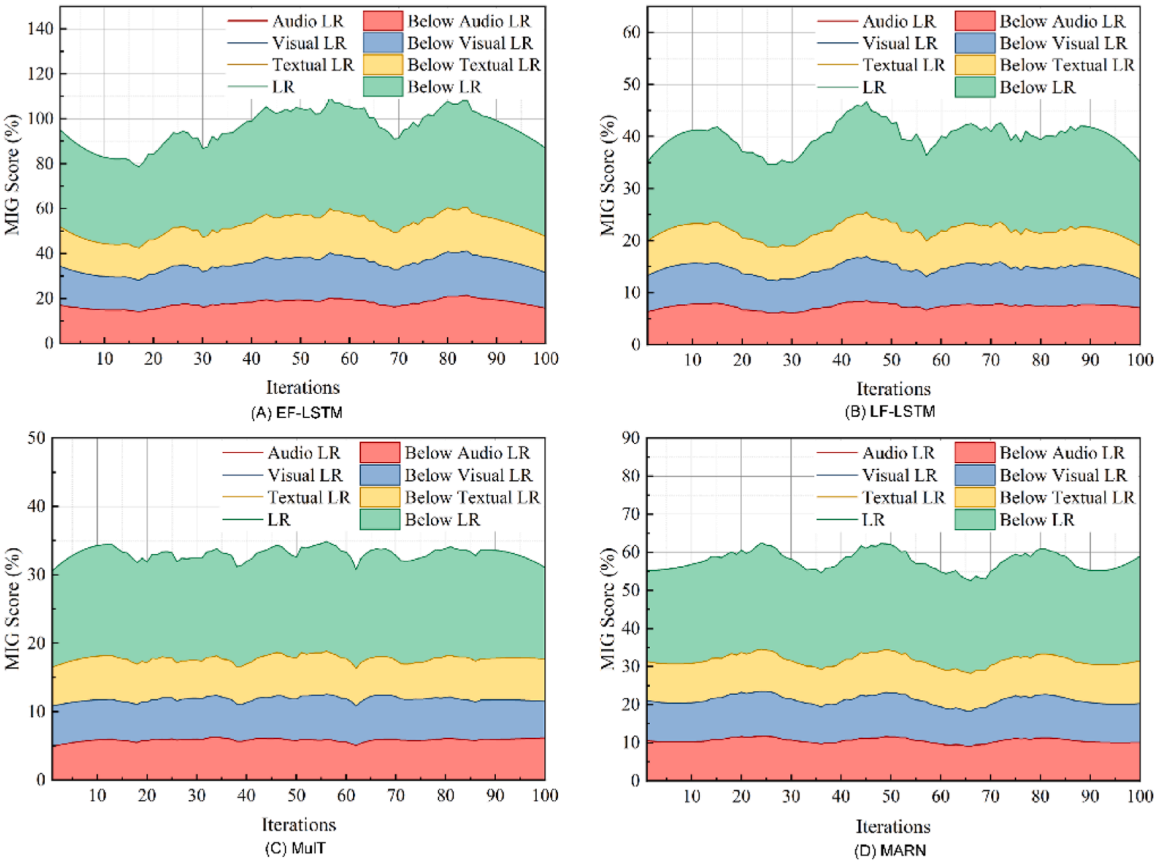
(a) CMU-MOSEI

(b) CMU-MOSI

Fig. 7. Comparison of classification results between disentangled latent representations and original unimodal representations on the benchmark datasets. The DVR/DAR/DTR represent disentangled audio/video/textual representations, respectively. On the other hand, OVR/ OAR/OTR denote original video/audio/textual representations, respectively. Notably, we normalize all metrics to a range of integers to facilitate the presentation of results.

can be estimated by utilizing the joint distribution. Therefore, we illustrate the slight effect of representation concatenation on the experimental results by quantifying the relative proportion of each modality-related latent vector within ground truth matrices (i.e. the joint representation). The results are shown in Fig. 8 and Fig. 9. Notably, different colored areas represent the MIG Score-based proportion between the individual latent variable (i.e. $\mathscr{Z}_a/\mathscr{Z}_v/\mathscr{Z}_t/\mathscr{Z}_{a,v,t}$) and the ground truth matrices (i.e. $\mathscr{X}'_{\mathscr{J}}$). The MIG scores of LF-LSTM, EF-LSTM, MulT, and MARN are denoted as A-D, respectively. Audio/-Visual/Textual LR represents $\mathscr{Z}_a/\mathscr{Z}_v/\mathscr{Z}_t$. LR represents $\mathscr{X}'_{\mathscr{J}}$. Notably, the MIG scores corresponding to each iteration are averaged over 100 epochs and all MIG Scores are smoothed by the signal-processing function.

Since current fusion algorithms are mainly based on the recurrent cell and attention mechanisms, we choose EF-LSTM, LF-LSTM, MulT, and MARN as examples of the recurrent cell-based method, the attention-based model, and the combination of the two, respectively. As visualization results presented in Fig. 8 and Fig. 9, it is clear that the MIG score ratios of disentangled latent variables are essentially equivalent to the sum of that of the modality-related latent variables, observing from the comparison of *Below LR* and the stacked area of *Below Audio/Visual/Textual LR*. From the visualization of the MIG Score on the CMU-MOSI



(A) EF-LSTM

(B) LF-LSTM

(C) MulT

(D) MARN

Fig. 8. Visualization of the MIG Score on the CMU-MOSI dataset.

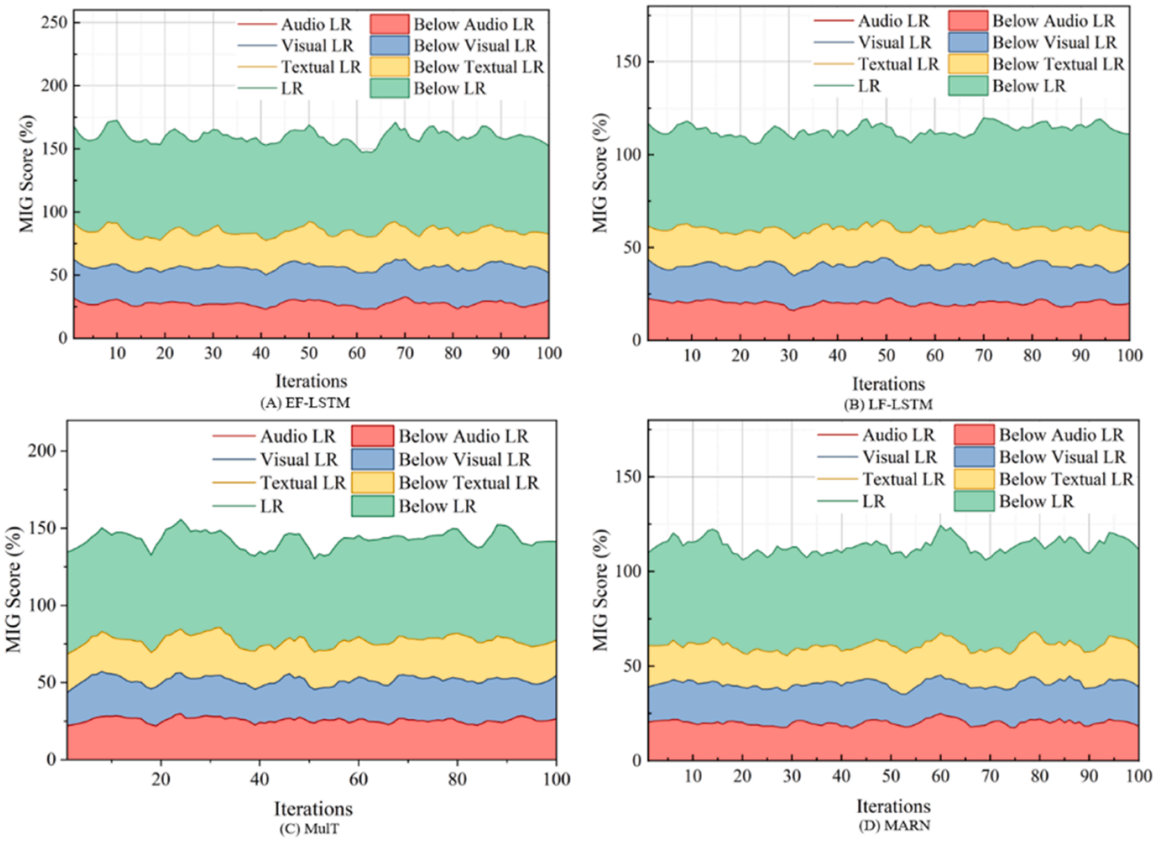


Fig. 9. Visualization of the MIG Score on the CMU-MOSEI dataset.

dataset, we note that the MIG scores of the recurrent cell-based multimodal fusion approach such as EF-LSTM outperform that of the attention-based approach across various fusion strategies. On the one hand, the recurrent unit-based methods are more suitable for analyzing temporal dependencies between unimodal features in multimodal learning tasks, decoupling joint representations, fused by LSTM-based methods, facilitates the acquisition of superior unimodal representations. On the other hand, attention mechanism-based methods prove more adept at probing correlations among cross-modal features. However, the condition that assumes the independence of individual modal vectors during the learning of disentangled representations might compromise the effectiveness of the attentional mechanism. This weakening effect contributes to distinctions in the MIG scores between the two methods. From the results shown in Fig. 9, the MIG scores of all methods are higher than the results on CMU-MOSI. This could be attributed to the CMU-MOSEI offering more comprehensive and high-dimensional features of each modality. Additionally, the comparison of methods based on the recurrent unit-based models demonstrates that feature-oriented fusion surpasses decision-oriented fusion in generating high-quality disentangled latent variables, as observed from the comparative results between EF-LSTM and LF-LSTM.

In terms of the proportion of independent modes, no matter what the dataset is, each independent mode contributes to the generation of a complementary and complete joint representation and the contribution of each modality is almost equal. This uniform contribution underscores the distributional constraint layer's effectiveness in resolving the modal bias issue. As for the proportion of concatenated representations, the MIG scores associated with the concatenation operation nearly equal the sum of the scores from the three independent modes, which indicates the concatenation operation solely establishes a structural connection among unimodal representations. This observation suggests that, in this experiment, the concatenation operation aims to facilitate the exploration of model input complementarity without interfering with the results of the model fusion performance evaluation in the subsequent model fusion

Effectiveness Analysis. The disentangled latent representations generated by the disentangled variational autoencoder encapsulate sentiment information from various independent modalities. To evaluate the multimodal information fusion capability of the sentiment analysis model, we compare the performance difference between the concatenated disentangled latent representations and the joint representations. The change in the performance degradation ratios γ(% %) reflects the model's ability to fuse multimodal information. The specific details of the experiments are as follows.

1. We conduct fusion performance evaluation experiments on the EF-LSTM, LF-LSTM, RMFN, LSTHM, MulT, MARN, RAVEN, and MFN models using the CMU-MOSEI and CMU-MOSEI datasets. The experimental results are shown in Tables 3 and 4, where $J \sim \mathscr{A}cc_*$(%) and $D \sim \mathscr{A}cc_*$(%) denote the accuracy results based on the joint representation and the corresponding tandem decoupled implicit representation, respectively.
2. To explore the information fusion capability of multimodal sentiment analysis models across different sentiment types, we conducted a single classification comparison experiment based on the data structure attributes of the IEMOCAP dataset. This experiment compared the performance of the EF-LSTM, LF-LSTM, RMFN, LSTHM, MulT, MARN, RAVEN, and MFN models on individual sentiment categories. The results are presented in Table 6, where $S \sim *$ and $DS \sim *$ denote the performance based on independent input modal features and the corresponding independent modal decoupled implicit representation, respectively.
3. To investigate the information fusion capability of multimodal sentiment analysis models across various language datasets, we

Table 3
The Accuracy Degradation Ratios of Methods on CMU-MOSEI Datasets.

Type	Dataset	CMU-MOSEI					
	Method	$J\sim\mathscr{A}cc_2$(%)	$D\sim\mathscr{A}cc_2$(%)	γ(%)	$J\sim\mathscr{A}cc_7$	$D\sim\mathscr{A}cc_7$(%)	γ(%)
Rubm	EF-LSTM	69.07	70.98	2.765	45.34	41.37	8.756
Rubm	LF-LSTM	66.42	70.96	6.835	44.76	41.37	5.945
Rubm	RMFN	78.61	71.05	9.617	46.16	41.37	10.38
Rubm	LSTHM	64.81	70.90	9.350	45.55	41.37	9.177
Abm	MulT	71.06	70.34	**1.013**	42.21	41.37	**1.990**
Abm	MARN	77.80	71.05	8.676	46.00	41.37	10.07
Abm	RAVEN	77.13	70.96	7.999	47.83	41.37	13.51
Rubm*	MFN	77.23	70.83	8.286	48.30	41.37	14.35

Table 4
The Accuracy Degradation Ratios of Methods on CMU-MOSI Datasets.

Type	Dataset	CMU-MOSI					
	Method	$J\sim\mathscr{A}cc_2$(%)	$D\sim\mathscr{A}cc_2$(%)	γ(%)	$J\sim\mathscr{A}cc_7$	$D\sim\mathscr{A}cc_7$(%)	γ(%)
Rubm	EF-LSTM	73.03	46.50	36.33	30.76	17.34	43.63
Rubm	LF-LSTM	72.89	46.36	36.39	30.03	16.34	45.52
Rubm	RMFN	73.18	44.90	38.64	30.90	18.22	41.03
Rubm	LSTHM	70.41	44.17	37.27	26.38	15.45	41.43
Abm	MulT	60.50	44.75	26.03	24.49	15.45	36.91
Abm	MARN	72.01	44.90	37.65	31.63	14.43	54.38
Abm	RAVEN	74.64	43.88	41.25	32.22	13.84	57.05
Rubm*	MFN	72.30	45.77	36.69	30.90	15.45	50.00

additionally compared the performance of the LF-DNN, MCTN, LMF, and MISA models on the CH-SIMS dataset. The experimental results are presented in Table 7, where $J\sim\mathscr{A}cc_*$(%) and $D\sim\mathscr{A}cc_*$(%) denote the accuracy results based on the joint representation and the corresponding tandem decoupled implicit representation, respectively.

From the accuracy shown in Table 3, we observe a notable decline in experimental performance when employing disentangled latent representations compared to predictions based on joint representations in the CMU-MOSI dataset, but the classification results in the CMU-MOSEI dataset remain essentially unchanged. The MulT represents the exhibits minimum degradation ratio among all comparison algorithms in benchmark datasets. Since the multimodal baselines exhibit large variations in making classification predictions with latent and joint representations on the CMU-MOSI dataset, we conduct a randomized experiment on this dataset and calculate the degradation ratios for all metrics. Table 4 illustrates the relative degradation ratios of multimodal fusion approaches between the joint representations and disentangled latent representations. Particularly, we found that the MulT model has much less performance gap in accuracy ($\mathscr{A}cc_2 = 29.94$ % and $\mathscr{A}cc_7 = 36.52$ %) and error ($\mathscr{M}ae = 11.05$ %) metrics than other methods, indicating the MulT exhibits a relatively smaller network degradation.

Table 5 reveals that models relying on a single mechanism consistently exhibit superior performance across all metrics. For instance, in the CMU-MOSI dataset, both the recurrent cell-based (i.e., EF-LSTM and LF-LSTM) and attention-based models (i.e., MulT) outperform the combined model, as evident in the degradation ratios of the class of methods. This discrepancy may stem from the optimization problem of multi-mechanism loss functions. Specifically, the LSTM-based method excels in addressing temporal feature-dependent issues related to modal features, while the attention-based mechanism is more adept at extracting strongly correlated cross-modal features. When multiple mechanisms are trained in parallel in the combined model, balancing the different error losses and parameter updates to achieve optimal results is more painstaking, thus leading to slight performance degradation.

Table 5
The Degradation Ratios of Each Performance Metric on the CMU-MOSEI Dataset.

Method	Type	$\gamma_{\mathscr{A}cc_2}$(%)	$\gamma_{\mathscr{A}cc_7}$(%)	$\gamma_{\mathscr{F}1}$(% %)	$\gamma_{\mathscr{M}ae}$(%)
EF-LSTM	Rubm	36.27	42.81	45.80	39.45
LF-LSTM	Rubm	35.94	42.53	**38.99**	38.63
RMFN	Rubm	39.13	39.02	54.52	54.16
LSTHM	Rubm	37.78	43.92	54.2	36.70
MulT	Abm	**29.94**	**36.52**	52.6	**11.05**
MARN	Abm	37.53	56.59	60.66	40.65
RAVEN	Abm	40.98	56.54	56.67	46.37
MFN	Rubm*	36.44	49.55	48.31	44.94

Additionally, the classification results of the multimodal sentiment analysis model across different sentiment types indicate a tendency of the model to categorize specific sentiment types more accurately. The results are presented in Table 6: the multimodal sentiment analysis model exhibits the lowest performance degradation ratio for the sentiment type labeled ’sad’ ($\mathscr{A}cc_2$(%): [0.83 %∼5.77 %], $\mathscr{F}1$(%): [1.28 % ∼12.96 %]), and the highest performance decay rate for the sentiment type labeled ’neutral’ ($\mathscr{A}cc_2$(%): 34.9 %, $\mathscr{F}1$(%): 65.27 %). Therefore, sentiment type significantly affects the information fusion ability of multimodal sentiment analysis models.

The results of the comparison experiments using different language datasets are shown in Table 7. The experimental results indicate that the multimodal sentiment analysis model experiences significant performance degradation across various language datasets. Specifically, the LF-DNN obtains the minimum performance degradation ratio in the CMU-MOSI dataset, i.e., $\mathscr{A}cc_2$(%): 27.67 %, $\mathscr{A}cc_7$(%): 55.93 %.The LF-DNN and MISA have the lowest performance decay rates in the CH-SIMS dataset, i.e., LF-SNN∼$\mathscr{A}cc_2$(%): 59.54 %, MISA∼$\mathscr{A}cc_5$(%): 59.88, respectively. Furthermore, the trend of the experimental results reveals that as the complexity of the model structure increases (LF-DNN → LMF → MISA), the performance degradation ratio of the multimodal sentiment analysis model also gradually increases.

Analysis of feature distribution. We introduced the evaluating metrics (MIG score) to quantify the similarity between the latent and joint variables. Taking the LF-LSTM model as an example, we show the correlation visualizations between independent modality-related vectors (the disentangled key factors) and coupled joint matrices (the ground truth data), which can be seen in Fig. 10.

In terms of the overall correlation distribution shown in Fig. 10:

Table 6

Performance Degradation Rations of Different Models on the IEMOCAP Dataset for Classification of Individual Sentiment Types.

Method	Label	$S \sim \mathscr{A}cc_2$ (%)	$DS \sim \mathscr{A}cc_2$ (%)	γ(%)	$S \sim \mathscr{F}1$ (%)	$DS \sim \mathscr{F}1$ (%)	γ(%)
EF-LSTM	Neutral	69.51	55.07	20.77	69.38	39.12	43.61
	Happy	86.78	93.75	8.03	83.93	90.73	**8.10**
	Sad	84.97	80.07	**5.77**	84.11	71.21	15.34
	Angry	86.88	81.25	6.48	87.11	72.84	16.38
LF-LSTM	Neutral	40.83	55.08	34.90	23.67	39.12	65.27
	Happy	85.61	93.75	9.51	78.96	90.72	14.89
	Sad	79.42	80.08	**0.83**	70.31	71.21	**1.28**
	Angry	75.80	81.25	7.19	65.36	72.84	11.44
RMFN	Neutral	51.70	55.08	6.54	50.50	39.12	22.53
	Happy	85.61	93.75	9.51	78.96	90.72	14.89
	Sad	79.42	80.07	**0.82**	70.32	71.22	**1.28**
	Angry	75.80	81.25	7.19	65.37	72.84	11.43
LSTHM	Neutral	40.83	55.08	34.90	23.67	39.12	65.27
	Happy	85.61	93.75	9.51	78.96	90.72	14.89
	Sad	79.42	80.08	**0.83**	70.31	71.21	**1.28**
	Angry	75.80	81.25	7.19	65.36	72.84	11.44
MulT	Neutral	61.30	44.92	26.72	61.58	27.85	54.77
	Happy	86.35	93.75	8.57	82.97	90.73	9.35
	Sad	78.47	80.07	**2.04**	73.01	71.22	**2.45**
	Angry	82.41	81.25	1.41	81.36	72.84	10.47
MARN	Neutral	68.12	44.92	34.06	67.09	27.85	58.49
	Happy	85.93	93.75	9.10	82.29	90.72	10.24
	Sad	81.56	80.08	**1.81**	81.81	71.21	**12.96**
	Angry	86.35	81.25	5.91	85.98	72.84	15.28
RAVEN	Neutral	64.61	44.92	30.48	64.32	27.84	56.72
	Happy	84.75	93.75	10.62	81.04	90.72	11.94
	Sad	82.08	80.07	**2.45**	80.78	71.22	**11.83**
	Angry	81.34	81.25	0.11	78.81	72.84	7.58
MFN	Neutral	44.88	55.08	22.73	37.82	39.12	3.44
	Happy	85.60	93.75	9.52	78.96	90.72	14.89
	Sad	79.42	80.07	**0.82**	70.31	71.22	**1.29**
	Angry	75.80	81.25	7.19	65.37	72.84	11.43

Table 7

The Performance Degradation Ratios of Complex Models on CMU-MOSI and CH-SIMS Datasets.

Dataset	CMU-MOSI					
Method	$J \sim \mathscr{A}cc_2$(%)	$D \sim \mathscr{A}cc_2$(%)	γ(%)	$J \sim \mathscr{A}cc_7$(%)	$D \sim \mathscr{A}cc_7$(%)	γ(%)
LF-DNN	76.39	55.25	**27.67**	35.06	15.45	**55.93**
LMF	77.55	44.75	42.30	41.69	15.45	62.94
MISA	77.84	44.75	42.51	35.28	15.45	56.21
Dataset	CH-SIMS					
Method	$J \sim \mathscr{A}cc_2$(%)	$D \sim \mathscr{A}cc_2$(%)	γ(%)	$J \sim \mathscr{A}cc_5$(%)	$D \sim \mathscr{A}cc_5$(%)	γ(%)
LF-DNN	75.71	30.63	**59.54**	41.58	15.1	63.68
LMF	73.96	30.63	58.59	38.29	15.1	60.56
MISA	78.77	30.63	61.11	37.64	15.1	**59.88**

Correlation Distribution (All Latent Representation), we can see that the latent variables of the textual modality exhibit a relatively uniform correlation distribution, in contrast to the other modes that display a noticeably sparse correlation distribution. An explicable rationale lies in the context-dependent nature of textual modality, which fosters semantic correlation among textual features across various periods. This characteristic enables a substantial contribution of most representations to downstream tasks. In contrast, other modalities determine the sentiment polarity depending on a salient representation.

Another significant distribution characteristic observed from the heat maps is that the key factors contributing to the joint representation are more concentrated in the middle and latter parts of the modality-related latent representations, i.e. the dark blue rectangles indicating high correlations are more densely distributed in the middle and the right side of the figure. This was probably caused by the emotional cumulative effect, which gradually increased the emotional intensity over time. Thus, some concluding points with strong emotional tendencies usually appear at the end of the dialogue or monologue. An interesting phenomenon can be found in the correlation distribution visualization of independent modalities, as shown in Fig. 10: Correlation Distribution of textual, visual, and audio latent representations. We found that the complementary information distribution between different modalities in multimodal fusion was the key to improving the performance of multimodal fusion strategies in downstream tasks. The positions of the dark blue rectangles in different modalities heat-maps are complimentary, indicating that the latent representations representing sentiment information among different modalities are complementary.

We offer detailed visual representations that highlight the issue of imbalanced distribution characteristics in joint representations. Specifically, we depict the distributions of ground truth labels, network predictions, joint representations, and unimodal latent representations using feature scatter plots. While each mode generates complementary feature distributions with an equal number of generated feature points, the feature coupling in multimodal fusion results in a significant reduction in the number of feature points within certain modal ranges in the joint characterization. These can be observed from the comparison of joint features shown in Fig. 11(A) and each unimodal feature shown in Fig. 11(D)-(F). The values of the joint representation exhibit higher density within the range of [0.4, 1.0] and lower density within the range of [0, 0.4], according to the distribution results shown in Fig. 10(A),

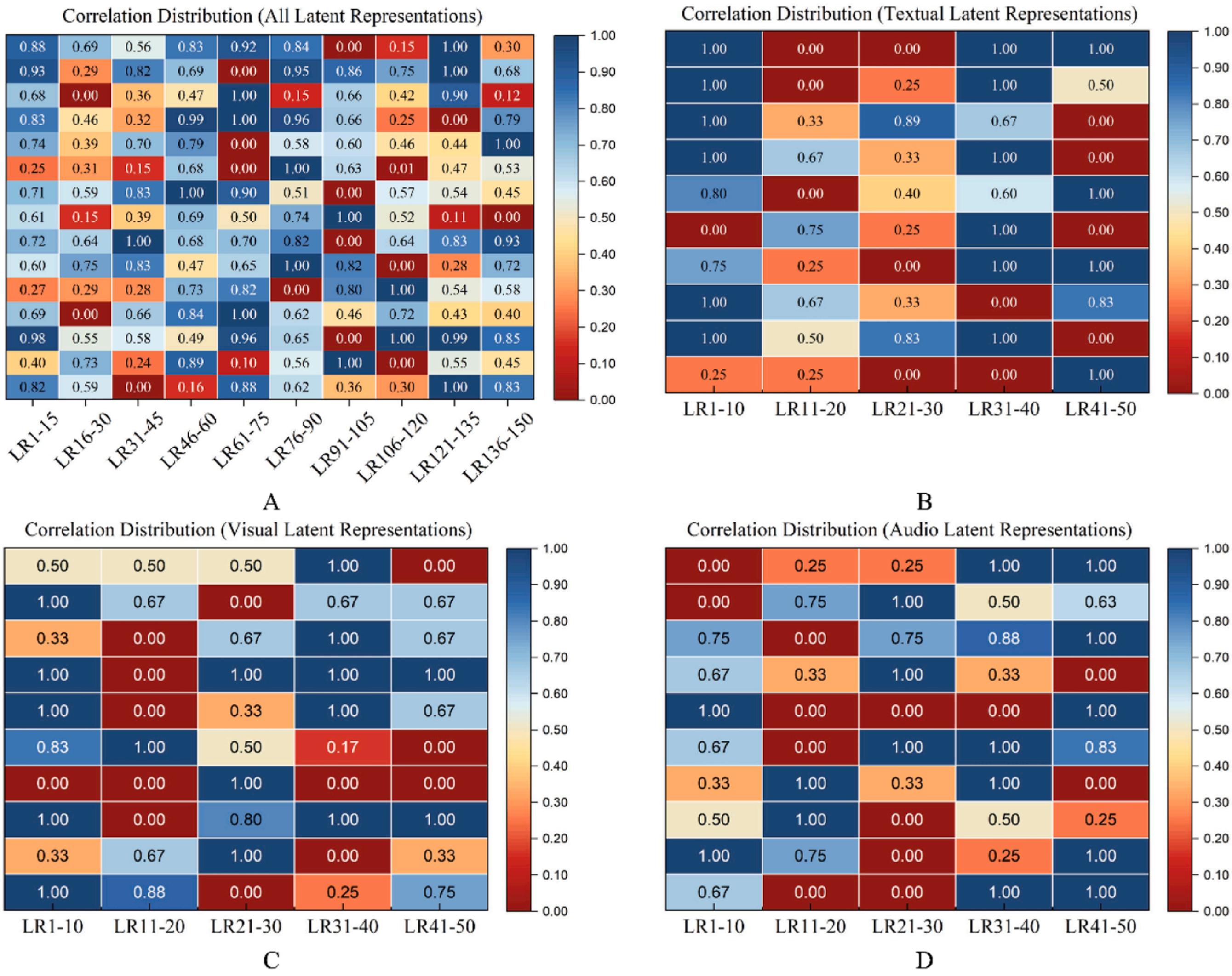


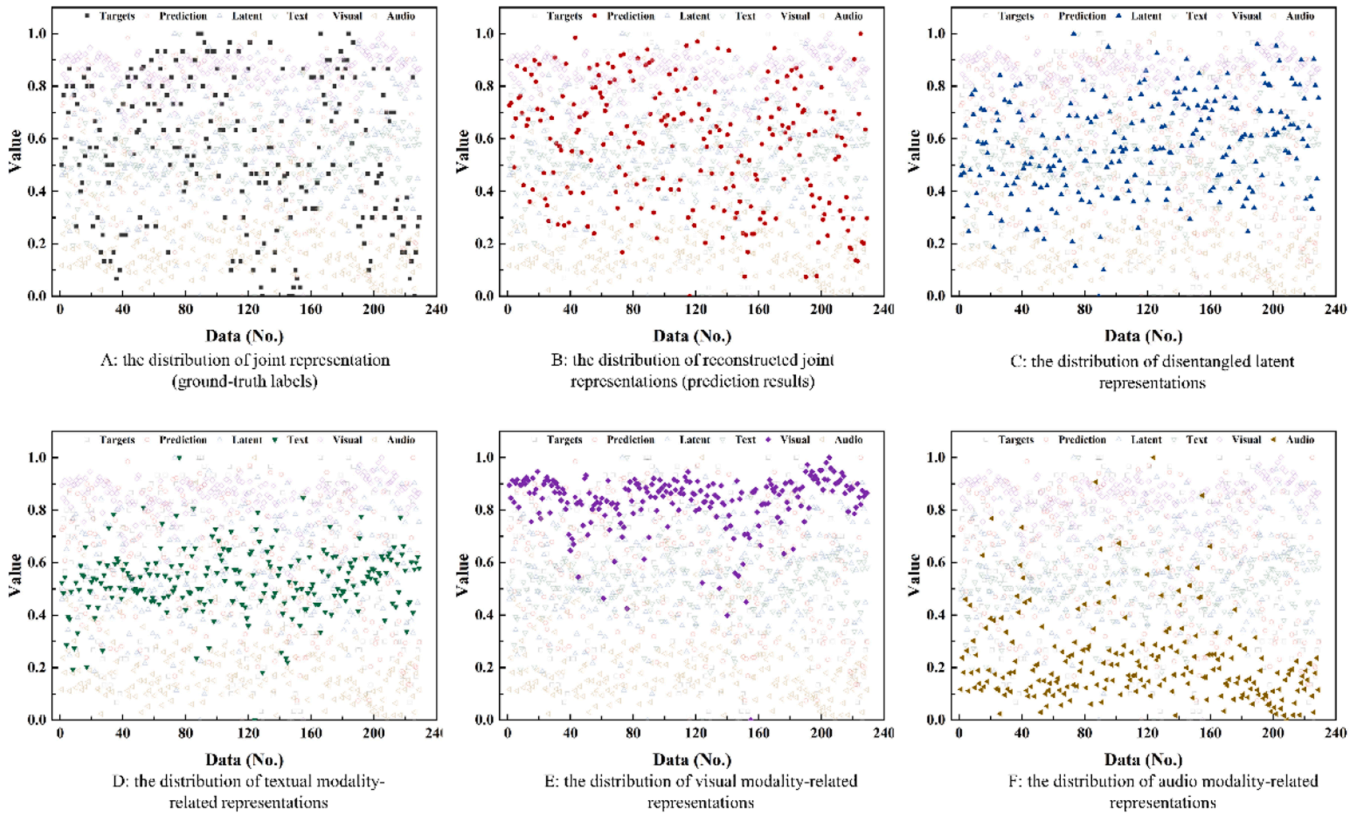
Fig. 10. Correlation Distribution between LF-LSTM Latent and Joint Representations. The index of latent representation is incremented from top to bottom. Different index ranges indicate different modality representations in correlation distribution (all latent representations), i.e., LR1–50: Textual modality, LR51–100: visual modality, LR101–150: audio modality.

which indicates that the text and visual modalities exhibit a higher contribution to the joint representation compared to speech modalities. From the phenomenon, we can conclude that the contribution of different modalities to the joint representation is imbalance (i.e. the modality imbalance caused by feature coupling), which provides a comprehensive explanation for why some unimodal-based sentiment analysis models (e.g., text-based modality) outperform multimodal fusion approaches in MSA.

Correlation coefficient analysis: We briefly examine the crucial factors that could influence multimodal fusion in terms of data correlation and statistical distribution. Firstly, we find that the strength of the correlation between the joint representation and the ground labels is not a conclusive factor in determining the effectiveness of model fusion performance. As evident from the classification accuracy metrics presented in Table 1, both joint and reconstructed joint representations exhibit similar performance in the MSA task. However, notable disparities are observed in Pearson's correlation metric between the two representations. To provide a more detailed illustration of the relationship between the statistical distributions of various representations and data labels, we visualized the statistical distributions of these representations. The results are presented in Fig. 12. All values depicted in the figure have been normalized.

While the joint representation distributions of certain models may align more closely with the true label distributions, their classification accuracies are not exceptional, as evidenced by the EF-LSTM and LF-LSTM models in Fig 12. Upon comparing the label distributions of the two datasets, it is observed that the CMU-MOSI data demonstrates a multi-peak distribution, as illustrated in Fig. 12(b) −2 (label distribution). This observation might explain the sharp decay in accuracy shown in Table 3 (CMU-MOSI dataset). The multi-peak distribution frequently results in complex decision boundaries. During network training, it is difficult for the multimodal fusion model to update parameters and generalize patterns based on data labels. In particular, when the multimodal features are unbalanced it will lead to the prediction results being biased towards a certain peak and result in lower model recall.

For multimodal fusion models, those relying on a single mechanism are more straightforward to train and generalize compared to models based on the combination of multiple mechanisms, but the latter exhibits a higher performance ceiling than the former. Therefore, striking a balance between the two is crucial for achieving concise and efficient multimodal models. In addition, there is no necessity for generating representations that align closely with the distribution of data labels to boost the classification performance of multimodal models, and ground labels with unimodal distributions are more useful for improving the performance of multimodal models



A: the distribution of joint representation (ground-truth labels)

B: the distribution of reconstructed joint representations (prediction results)

C: the distribution of disentangled latent representations

D: the distribution of textual modality-related representations

E: the distribution of visual modality-related representations

F: the distribution of audio modality-related representations

Fig. 11. Unbalanced distribution of modal features. Prediction: distribution of the LF-LSTM model. Target: data label distribution. Latent: distribution of latent representations generated by the distribution constraint layer. Visual/Audio/Text: distribution of visual/audio/textual modalities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

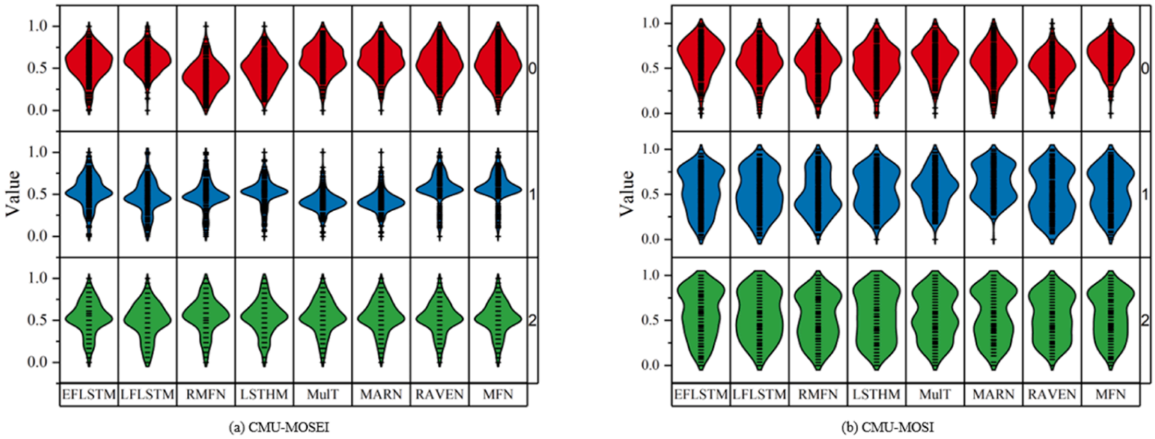


Fig. 12. Violin diagrams for statistical distributions. The values 0, 1, and 2 represent the reconstructed joint representations, fused joint representations, and ground labels, respectively.

6. Conclusion and future research

To comprehensively evaluate the fusion performance of state-of-the-art multimodal fusion methods and investigate the key factors that contribute to facilitating multimodal fusion in Multimodal Sentiment Analysis, this paper presented a Disentanglement-based Variable Auto-Encoder (DVAE). Specifically, the distribution constraint component was first proposed to generate modality-related latent representations by decoupling multimodal joint representation. Then, the modified combined loss term was designed in DVAE to facilitate the optimization of neural network weights and parameters by integrating inductive bias, signal reconstruction, and distribution constraint items. With the provided evaluation method, we can evaluate the fusion performance of multimodal models by contrasting the classification degradation ratios obtained from disentangled latent representations and the joint representation. The results from the disentanglement evaluation experiments confirm that DVAE can effectively facilitate the decoupling and reconfiguration of the joint representation without significantly compromising the model performance. The performance evaluation results on the CMU-MOSI and the CMU-MOSEI benchmark datasets indicate the proposed method can serve as an effective assessment tool for evaluating the information fusion capability of multimodal sentiment analysis models. Compared with 8 state-of-the-art methods employing different fusion mechanisms, it becomes evident that the distributional consistency between the fused features and the ground labels alone cannot determine the fusion effect of the multimodal model. Instead, the equalization effect among various fusion mechanisms in multimodal sentiment analysis, along with the single-peak characteristic of the ground label distribution, proves to be crucial in multimodal data fusion.

Our study introduces a novel methodology in disentangled representation learning, particularly in crafting disentangled representation learning models with generative constraints. This contributes to advancing the theoretical understanding of effectively segregating and rearranging various modal representations, offering insights into potential mechanisms for multimodal data fusion. Additionally, the disentangled variational auto-encoder model offers a pathway for practitioners to build enhanced and precise multimodal sentiment analysis systems. This advancement holds the potential to significantly enhance the performance of applications across domains such as social media analysis, customer feedback systems, and sentiment detection in multimedia content.

While DVAE demonstrates excellence in evaluating multimodal model performance and achieving feature decoupling, it demands substantial hardware resources to support model training, particularly when handling large-scale multimodal datasets. This is primarily attributed to the resource-intensive nature of training dedicated encoders for different modalities to achieve feature decoupling and generate hidden variables, which imposes a demand on memory resources. Our future work will be devoted to model lightweight and exploring how the proposed approach can be used for real-time content understanding generation and transformation between speech and sign language modalities, as well as deployed in other practical applications of interpretable AI.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

CRediT authorship contribution statement

Rongfei Chen: Writing – original draft, Software, Methodology, Investigation, Data curation. **Wenju Zhou:** Writing – review & editing, Supervision, Project administration. **Huosheng Hu:** Writing – review & editing, Supervision, Resources. **Zixiang Fei:** Writing – review & editing, Funding acquisition, Data curation. **Minrui Fei:** Writing – review & editing. **Hao Zhou:** Writing – review & editing, Resources.

Declaration of competing interest

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was Sponsored by the Shanghai Pujiang Program (No. 22PJ1403800) the Natural Science Foundation of China (NSFC) under Grant 62203290, and the China Scholarship Council (Grant Number 202106890006).

References

[1] H. Sun, H. Wang, J. Liu, Y.W. Chen, L. Lin, CubeMLP: an MLP-based model for multimodal sentiment analysis and depression estimation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3722–3729.

[2] Y. Ying, T. Yang, H. Zhou, Multimodal fusion for Alzheimer's disease recognition, Appl. Intell. 53 (2023) 16029–16040.

[3] R.Bhalla Pooja, A review paper on the role of sentiment analysis in quality education, SN Comput. Sci. 3 (2022) 469.

[4] L.P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, in: Proceedings of the 13th International Conference on Multimodal Interfaces, 2011, pp. 169–176.

[5] T. Winterbottom, S. Xiao, A. McLean, N.A. Moubayed, On modality bias in the tvqa dataset, arXiv Preprint (2020).

[6] X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via on-the-fly gradient modulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.

[7] S. Yan, D. Huang, M. Soleymani, Mitigating biases in multimodal personality assessment, in: Proceedings of the 2020 International Conference on Multimodal Interaction, 2020, pp. 361–369.

[8] Y. Deng, J. Ma, ReDFeat: recoupling detection and description for multimodal feature learning, IEEE Trans. Image Process. 32 (2022) 591–602.

[9] R. Das, T.D. Singh, Multimodal sentiment analysis: a survey of methods, trends, and challenges, ACM Comput. Surv. 55 (2023) 1–38.

[10] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106.

[11] M. Böhle, F. Eitel, M. Weygandt, K. Ritter, Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification, Front. Aging Neurosci. 11 (2019) 456892.

[12] S. Mandloi, M. Zuber, R.K. Gupta, An explainable brain tumor detection and classification model using deep learning and layer-wise relevance propagation, Multimed. Tools Appl. 83 (2024) 33753–33783.

[13] X. Xue, C. Zhang, Z. Niu, X. Wu, Multi-level attention map network for multimodal sentiment analysis, IEEE Trans. Knowl. Data Eng. 35 (2022) 5105–5118.

[14] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[15] M. Staudte, M.W. Crocker, Investigating joint attention mechanisms through spoken human–robot interaction, Cognition 120 (2011) 268–291.

[16] J. Huang, Y. Jiao, X. Liao, J. Liu, Z. Yu, Deep dimension reduction for supervised representation learning, IEEE Trans. Inf. Theory 70 (2024) 3583–3598.

[17] X. Wang, H. Chen, W. Zhu, Disentangled representation learning for multimedia, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9702–9704.

[18] S. Van Steenkiste, F. Locatello, J. Schmidhuber, O. Bachem, Are disentangled representations helpful for abstract visual reasoning? Adv. Neural Inf. Process. Syst. 32 (2019) 14245–14258.

[19] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, others, A systematic review on affective computing: emotion models, databases, and recent advances, Inf. Fusion 83 (2022) 19–52.

[20] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: a survey, Inf. Fusion 95 (2023) 306–325.

[21] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.

[22] A. Zadeh, P. Vij, P.P. Liang, E. Cambria, S. Poria, L.P. Morency, Multi-attention recurrent network for human communication comprehension, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence 2018, AAAI, 2018, pp. 5642–5649.

[23] M. Hou, Z. Zhang, C. Liu, G. Lu, Semantic alignment network for multi-modal emotion recognition, IEEE Trans. Circuits Syst.Video Technol. 33 (2023) 5318–5329.

[24] F. Locatello, S. Bauer, M. Lucie, G. Rätsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019 2019-June, 2019, pp. 7247–7283.

[25] G. Yi, C. Fan, K. Zhu, Z. Lv, S. Liang, Z. Wen, G. Pei, T. Li, J. Tao, Vlp2msa: expanding vision-language pre-training to multimodal sentiment analysis, Knowl. Based Syst. 283 (2024) 111136.

[26] J. Li, X. Zhang, F. Li, S. Duan, L. Huang, Acoustic-articulatory emotion recognition using multiple features and parameter-optimized cascaded deep learning network, Knowl. Based Syst. 284 (2024) 111276.

[27] C. Cai, Y. He, L. Sun, Z. Lian, B. Liu, J. Tao, M. Xu, K. Wang, Multimodal Sentiment Analysis based on Recurrent Neural Network and Multimodal Attention, in: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, ACM, Virtual Event China, 2021, pp. 61–67.

[28] X. Li, J. Liu, Y. Xie, P. Gong, X. Zhang, H. He, Magdra: a multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition, Knowl. Based Syst. 283 (2024) 111126.

[29] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, X. Huang, TeFNA: text-centered fusion network with crossmodal attention for multimodal sentiment analysis, Knowl. Based Syst. 269 (2023) 110502.

[30] R. Wadawadagi, V. Pagi, Sentiment analysis with deep neural networks: comparative study and performance assessment, Artif. Intell. Rev. 53 (2020) 6155–6195.

[31] D. Zimbra, A. Abbasi, D. Zeng, H. Chen, The state-of-the-art in Twitter sentiment analysis: a review and benchmark evaluation, ACM TMIS 9 (2018) 1–29.

[32] S.D. Pande, B.R. Altahan, S.H. Ahammad, A.S. Mane, S. Inthiyaz, L.K. Smirani, M. A. Hossain, A.N.Z. Rashed, Assessment and recommendation of neural networks and precise techniques for sentiment systems analysis, J. Ambient Intell. Humaniz. Comput. 14 (2023) 11285–11299.

[33] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821.

[34] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37.

[35] C.S. Chan, H. Kong, G. Liang, A comparative study of faithfulness metrics for model interpretability methods, arXiv Preprint (2022).

[36] Q. Zhao, J. Liu, Z. Kang, Z. Zhou, TraceNet: tracing and locating the key elements in sentiment analysis, Knowl. Based Syst. 277 (2023) 110792.

[37] R. Huang, Q. Chen, J. Tang, J. Song, The Influence of Word Embeddings on the Performance of Sentiment Classification, Int. J. Comput. Inf. Technol. 4 (2023), 1–1.

[38] A. Feng, Z. Chen, S. Zhou, X. Wu, Embeddings and convolution, is that the best you can do with sentiment features?, in: Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN) IEEE, 2019, pp. 1–8.

[39] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, A. Lerchner, Towards a definition of disentangled representations, arXiv Preprint (2018).

[40] D. Yang, S. Huang, H. Kuang, Y. Du, L. Zhang, Disentangled representation learning for multimodal emotion recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1642–1651.

[41] M. Tschannen, O. Bachem, M. Lucic, Recent advances in autoencoder-based representation learning, arXiv Preprint (2018).

[42] I. Daunhawer, T.M. Sutter, R. Marcinkevičs, J.E. Vogt, Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models, in: Proceedings of the Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, 2020, pp. 459–473.

[43] M. Cranmer, C. Cui, D.B. Fielding, S. Ho, A. Sanchez-Gonzalez, K. Stachenfeld, T. Pfaff, J. Godwin, P. Battaglia, D. Kochkov, Disentangled sparsity networks for explainable AI, in: Proceedings of the Workshop on Sparse Neural Networks, 2021.

[44] K. Laenen, M.F. Moens, Learning explainable disentangled representations of e-commerce data by aligning their visual and textual attributes, Computers 11 (2022) 182.

[45] M. Ju, W. Song, S. Sun, Y. Ye, Y. Fan, S. Hou, K. Loparo, L. Zhao, Dr. emotion: disentangled representation learning for emotion analysis on social media to improve community resilience in the COVID-19 era and beyond, in: Proceedings of the Web Conference 2021, 2021, pp. 518–528.

[46] Y. Zhang, Y. Zhang, W. Guo, X. Cai, X. Yuan, Learning disentangled representation for multimodal cross-domain sentiment analysis, IEEE Trans. Neural Netw. Learn Syst. 34 (2022) 7956–7966.

[47] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

[48] A. Zadeh, R. Zellers, E. Pincus, L.P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv Preprint (2016).

[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] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, Ch-Sims: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3718–3727.

[51] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, D. Song, What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis, Information Fusion 66 (2021) 184–197.

[52] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, NIH Public Access, 2019, pp. 6558–6569.

[53] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.P. Morency, Words can shift: dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 7216–7223.

[54] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.

[55] D. Ghosal, M.S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, P. Bhattacharyya, Contextual inter-modal attention for multi-modal sentiment analysis, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3454–3466.

[56] A. Zadeh, S. Poria, P.P. Liang, E. Cambria, N. Mazumder, L.P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence 2018, AAAI, 2018, pp. 5634–5641.

[57] J. Williams, S. Kleinegesse, R. Comanescu, O. Radu, Recognizing emotions in video using multimodal DNN feature fusion, in: Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), 2018, pp. 11–19.

[58] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, arXiv Preprint (2018).

[59] D. Hazarika, R. Zimmermann, S. Poria, Misa: modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[60] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7482–7491.

[61] T.Q. Chen, X. Li, R. Grosse, D. Duvenaud, Isolating sources of disentanglement in variational autoencoders, in: Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, 2018, pp. 2615–2625.