MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis

Ronghao Lin and **Haifeng Hu**[*]
Sun Yat-sen University, China
linrh7@mail2.sysu.edu.cn, huhaif@mail.sysu.edu.cn

Abstract

When applying multimodal machine learning in downstream inference, both joint and coordinated multimodal representations rely on the complete presence of modalities as in training. However, modal-incomplete data, where certain modalities are missing, greatly reduces performance in Multimodal Sentiment Analysis (MSA) due to varying input forms and semantic information deficiencies. This limits the applicability of the predominant MSA methods in the real world, where the completeness of multimodal data is uncertain and variable. The generation-based methods attempt to generate the missing modality, yet they require complex hierarchical architecture with huge computational costs and struggle with the representation gaps across different modalities. Diversely, we propose a novel representation learning approach named MissModal, devoting to increasing robustness to missing modality in a classification approach. Specifically, we adopt constraints with geometric contrastive loss, distribution distance loss, and sentiment semantic loss to align the representations of modal-missing and modal-complete data, without impacting the sentiment inference for the complete modalities. Furthermore, we do not demand any changes in the multimodal fusion stage, highlighting the generality of our method in other multimodal learning systems. Extensive experiments demonstrate that the proposed method achieves superior performance with minimal computational costs in various missing modalities scenarios (flexibility), including severely missing modality (efficiency) on two public MSA datasets.

1 Introduction

With the proliferation of the Internet and the surge of user-generated videos, Multimodal Sentiment Analysis (MSA) has become an important and challenging research task that focuses on predicting sentiment with multiple modalities including text, audio, and vision (Morency et al., 2011; Poria et al., 2020). The previous models (Zadeh et al., 2017; Tsai et al., 2019a; Wang et al., 2019; Han et al., 2021) aim at learning a mapping function to fuse the information of different modalities and obtain distinguishable multimodal representations for sentiment inference. As shown in Figure 1, these MSA methods input utterances with multiple modalities to train the mapping function of multimodal representation in the supervised of ground truth labels, and apply the learned MSA models in the downstream testing to predict the sentiment of other utterances.

However, both training and testing pipelines in these MSA methods require complete-modal data, indicating the sensitivity to missing modalities for the mapping function. Missing any modality in testing causes differences in the distribution of input data from training, leading to performance drops of the mapping function. Due to the uncertainty and various modality settings in the real world, the demand for the integrity of modalities limits the application of the previous strategies of multimodal representation learning.

To deal with the issues of missing modalities, generation-based research emerges which focuses on leveraging the remained modalities to generate the missing modalities (Tsai et al., 2019b; Pham et al., 2019; Tang et al., 2021). These generative models have complex hierarchical architecture, which requires redundant training parameters and high computational costs in training. Besides, their generative performance is still challenged by the huge modality gap among different modalities, further limiting their application in the real world.

Different from the generative methods, we propose a novel multimodal representation learning approach named MissModal, devoted to increasing the model's robustness to missing modality

[*] Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1686–1702, 2023. https://doi.org/10.1162/tacl_a_00628
Action Editor: André F. T. Martins. Submission batch: 6/2023; Revision batch: 9/2023; Published 12/2023.

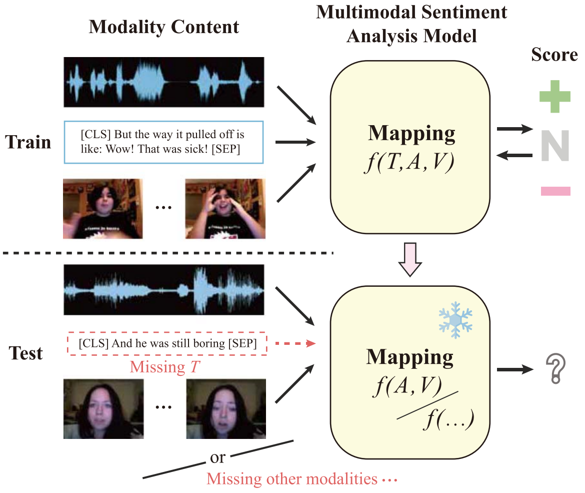

Modality Content
Multimodal Sentiment Analysis Model
Score
Train
[CLS] But the way it pulled off is like: Wow! That was sick! [SEP]
Mapping
f(T, A, V)
+
N
−
Test
[CLS] And he was still boring [SEP]
Missing T
Mapping
f(A, V)
f(...)
?
or
Missing other modalities ···

Figure 1: Illustration of missing modality in testing when applying the trained multimodal representation model in downstream application, where T, A, V denotes textual, acoustic, and visual modality, respectively.

in a classification way. Specifically, we utilize dependent modality-specific networks to learn representations for each modality. Then according to complete modalities—(text, audio, vision) and missing modalities (text), (audio), (vision), (text, audio), (text, vision), (audio, vision)—we adopt multimodal fusion networks with a consistent structure to learn the corresponding complete-modal and missing-modal representations. To transfer the semantic knowledge of complete modalities, we construct three constraints to align missing-modal and complete-modal representations, including geometric contrastive loss to utilize constrative learning at the level of samples, distribution distance loss to adjust the distribution of representations, and sentiment semantic loss to introduce supervise of sentiment labels.

Aiming at improving the downstream performance of MSA models in the real world, we retain the completeness of modalities in training, and then freeze the trained model for validation and testing with different missing rates for diverse modalities to evaluate the flexibility (randomly missing various modalities) and efficiency (severely missing modalities) of the proposed approach.

The contributions are summarized as follows:

1) We propose a novel multimodal representation learning approach named MissModal, devoted to increasing the robustness of MSA models to the issues of missing modalities in downstream applications.

2) Without generative methods, we construct three constraints to align the representations of missing and complete modalities, consisting of geometric contrastive loss, distribution distance loss, and sentiment semantic loss.

3) Extensive experiments on two publicly available MSA datasets with various settings of missing rates and missing modalities demonstrates the superiority of the proposed approach in both flexibility and efficiency.

2 Related Work

2.1 Multimodal Representation Learning

Diverse modalities such as natural language, motion videos, and vocal signals contain specific and complementary information on a common concept (Baltrušaitis et al., 2019). Multimodal representation learning focuses on exploring the intra- and inter-modal dynamics and learning distinguishable representations for various downstream tasks (Bugliarello et al., 2021). Recently, contrastive learning-based multimodal pre-trained models, *e.g.*, CLIP (Radford et al., 2021), WenLan (Huo et al., 2021), and UNIMO (Li et al., 2021), leverage contrastive learning to train transferrable mappings to bridge large-scale image-text pairs. The successful downstream application of these pre-trained models demonstrates the effectiveness of contrastive learning in aligning representations of different modalities.

As a task branch of multimodal machine learning, Multimodal Sentiment Analysis (MSA) aims at integrating the semantic information contained in different modalities, including textual, acoustic, and visual modalities, to predict the sentiment intensity of an utterance (Poria et al., 2020). Previous MSA methods mostly concentrate on designing effective multimodal fusion methods to explore the commonalities among different modalities (Zadeh et al., 2017; Rahman et al., 2020; Han et al., 2021) and learn informative multimodal representations. However, the training pipeline of explicit fusion strategies requires the presence of all modalities. Missing any modality in the downstream raises the differences of input condition between training and testing, causing wrong inference of sentiment in applications.

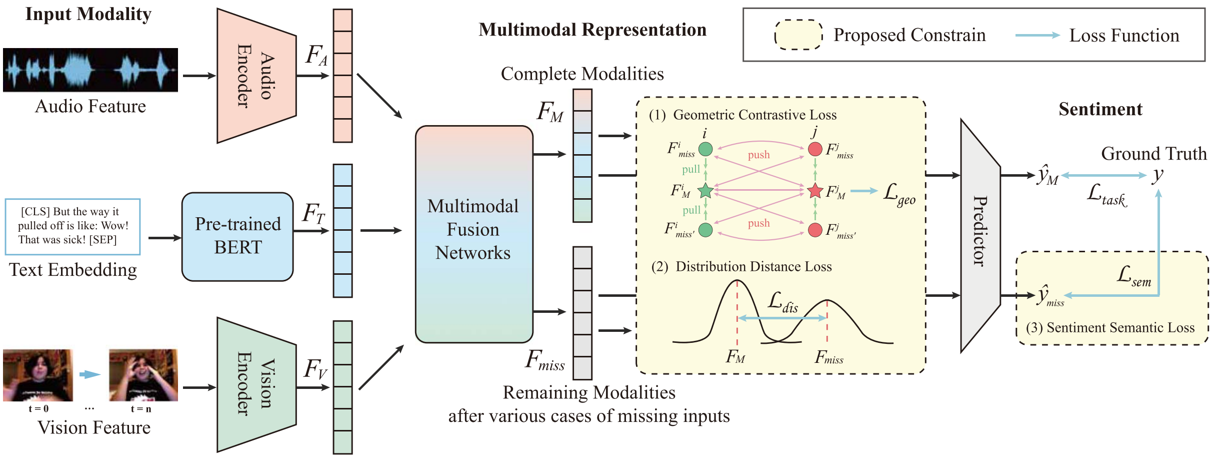


Figure 2: The overall architecture of the proposed MissModal. The missing-modal representations F_{miss} and complete-modal representations F_M are aligned with the guidance of the proposed losses $\mathcal{L}_{geo}$, $\mathcal{L}_{dis}$, and $\mathcal{L}_{sem}$ at both feature space and prediction level.

2.2 Missing Modality Issues

The aforementioned multimodal pre-trained models heavily depend on the completeness of modalities, making them fail to handle the issues of modality-incomplete data. As Ma et al. (2022) indicate, multimodal transformers (Hendreicks et al., 2021) are sensitive to missing modalities and the modality fusion strategies are dataset-dependent which significantly affects the robustness. Therefore, to address missing modality issues, generation-based methods (Ma et al., 2021; Vasco et al., 2022) are proposed to learn a prior distribution on modality-shared representation and infer the missing modalities in the modality-shared latent space, which are also employed in the MSA task (Tsai et al., 2019b; Pham et al., 2019; Tang et al., 2021). Nevertheless, these generation-based methods require large computational costs and the generative performance is limited by the huge modality gaps. Meanwhile, they mostly demand complex hierarchical model architecture which lack generality and efficiency in the downstream application. Differently from them, we are devoted to utilizing the classification approach instead of generation to reach the performance upper bound in the scenarios of missing modalities.

Recently, Hazarika et al. (2022) proposed robust training by utilizing missing and noisy textual input as data augmentation to train the state-of-the-art MSA models. However, the application of robust training is limited by the settings of single modality and fixed missing rates. Diversely, according to the missing rates and the diversity of missing modalities, we evaluate the performance by flexibility (randomly missing various modalities) and efficiency (severely missing modalities in testing) to show the improvement of robustness to missing modalities for the proposed approach.

3 Method

3.1 Task Definition

The input of MSA task is utterances which can be denoted as triplet (T, A, V), including textual modality $T \in \mathbb{R}^{\ell_T \times d_T}$, acoustic modality $A \in \mathbb{R}^{\ell_A \times d_A}$, and visual modality $V \in \mathbb{R}^{\ell_V \times d_V}$, where ℓ_U denotes the sequence length of corresponding modality and d_u denotes the feature dimension for $U \in \{T, A, V\}$. The goal is learning to map the multimodal data (T, A, V) into multimodal representations $F = f(T, A, V)$, where $F \in \mathbb{R}^{N_M \times d_M}$ can be utilized to infer the final sentiment scores $\hat{y} \in \mathbb{R}$. Specifically for better generalization performance in the downstream applications, the multimodal representation mapping f learned by the training data needs to handle the testing scenario as well, regardless of the completeness of modalities.

3.2 Model Architecture

To increase the robustness to missing modalities in testing, we propose a novel multimodal representation learning approach named MissModal, whose architecture is shown in Figure 2.

To obtain the modality-specific representations, we firstly adopt the pre-trained BERT (Devlin et al., 2019) to encode the input text embedding T and learn the textual representation, where the output embedding of the last Transformer layer is represented as:

$$F_t = BERT(T; \theta_T) \in \mathbb{R}^{\ell_T \times d_T} \tag{1}$$

Meanwhile, for the acoustic and visual modalities, we utilize two bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) to capture the temporal characteristics and two 3-layer unimodal Transformers (Vaswani et al., 2017) to further encode the global self-attention information. For $U \in \{A, V\}$, the audio and vision encoders are formulated as:

$$\begin{aligned} h_U &= bLSTM(U; \theta_U) \in \mathbb{R}^{\ell_U \times d_U} \\ F_U &= Transformer(h_U; \theta_U) \in \mathbb{R}^{\ell_U \times d_U} \end{aligned} \tag{2}$$

Specially, we take the [CLS] token of F_T and the embedding from the last time step of F_A and F_V, meaning that for $U \in \{T, A, V\}$, the modality-specific representations F_U satisfies $F_U \in \mathbb{R}^{d_U}$.

To capture the modality-shared dynamics, we utilize multimodal fusion networks to learn the latent interactions among different modalities. Specifically, devoted to better handling various cases of missing modality, we concatenate the modality-specific representations in seven ways to simulate seven input circumstances, including the settings of complete modalities, denoted as (T, A, V) and the remaining modalities after missing, denoted as $\{(T), (A), (V), (T, A), (T, V), (A, V)\}$. To highlight the effectiveness of MissModal, without losing generality, we adopt several simple MLPs with $Tanh$ activation layers as the fusion networks to extract the inter-modal information after concatenation, represented as:

$$\begin{aligned} F_M &= MLP([T; A; V]) \in \mathbb{R}^{d_M} \\ F_{miss} &= MLP([modality_i; ...]) \in \mathbb{R}^{d_M} \end{aligned} \tag{3}$$

where [;] denotes the concatenation of the modalities, F_M denotes the multimodal representation with complete modalities and F_{miss} denotes the representations with the inputs of remaining $modality_i$, $1 \leq i < 3$.

Note that the structure of multimodal fusion networks is optional and can be flexibly substituted by state-of-the-art multimodal fusion methods, illustrating the backward compatibility of the proposed approach.

3.3 Constraints when Missing Modalities

As shown in Figure 1, to improve the robustness of the model to the missing modalities, we propose three losses as constraints to align the missing-modal representations F_{miss} with the complete-modal ones F_M in the following.

3.3.1 Geometric Contrastive Loss

Poklukar et al. (2022) indicate that there are huge gaps between the modality-specific representations and complete representations leading to severe misalignment in the distribution space. Inspired by but different from Chen et al. (2020) and Poklukar et al. (2022), we introduce contrastive learning among the multimodal representations with complete modalities and the ones with different cases of missing modalities to geometrically align the representations from the same utterance samples in the supervision of sentiment labels.

Given a mini-batch of multimodal representations $\mathcal{B} = \{F_M^i, F_{miss}^i\}_{i=1}^B$, we define positive pairs as (F_M^i, F_{miss}^i) while the negative ones as (F_M^i, F_M^j), (F_{miss}^i, F_{miss}^j) and (F_M^i, F_{miss}^j) according to the ith and jth samples from the mini-batch $\mathcal{B}$. Then we compute the sum of similarities among the negative pairs as:

$$\begin{aligned} s_{p,q}(i,j) &= \exp(F_p^i \cdot F_q^j / \gamma), \;\; p, q \in \{M, miss\} \\ N_{p,q}(i) &= \sum_{i \neq j}^{B} s_{p,p}(i,j) + \sum_{j=1}^{B} s_{p,q}(i,j) \end{aligned} \tag{4}$$

where γ is a temperature hyperparameter regulating the probability distribution over distinct instances (Hinton et al., 2015). Similarly, the similarity of the positive pairs is denoted as $s_{p,q}(i,i)$, relating the missing-modal representations with the corresponding complete-modal ones.

By traversing all samples in the mini-batch, the geometric contrastive loss $\mathcal{L}_{geo}$ is represented as:

$$\mathcal{L}_{geo} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{s_{p,q}(i,i)}{N_{p,q}(i)} \tag{5}$$

The contrastive learning encourages multimodal fusion networks to transfer complete-modal information to the missing-modal representations, making them more distinguishable when handling the missing modalities issues in applications.

3.3.2 Distribution Distance Loss

To further enhance the similarity of F^i_{miss} and the corresponding F^i_M, we add L2 distance constraints to reduce the distribution distance among the missing-modal and complete-modal representations from the same sample. The distribution distance loss $\mathcal{L}_{dis}$ is represented as:

$$\mathcal{L}_{dis} = \frac{1}{B}\sum_{i=1}^{B} \|F^i_M - F^i_{miss}\|^2_2 \tag{6}$$

Both geometric contrastive loss $\mathcal{L}_{geo}$ and distribution distance loss $\mathcal{L}_{dis}$ increase the model's robustness in the feature space when missing modalities.

3.3.3 Sentiment Semantic Loss

Due to various semantic information contained in diverse modalities, missing modalities may result in different sentiments of the same utterance. For the consistency in the inference of sentiment polarity, we introduce the sentiment semantic loss $\mathcal{L}_{sem}$ to utilize the ground truth labels y to supervise the sentiment prediction of the missing-modal representations in the label space, which is denoted as:

$$\mathcal{L}_{sem} = \frac{1}{B}\sum_{i=1}^{B} |y^i - \hat{y}^i_{miss}| \tag{7}$$

3.3.4 Optimization Objective

In the MSA task, after obtaining the sentiment prediction $\hat{y}^i_M$ with complete modalities, we apply Mean Absolute Error (MAE) loss to conduct the regre ssion of sentiment labels. Along with the ground truth labels y, the task loss is formulated as:

$$\mathcal{L}_{task} = \frac{1}{B}\sum_{i=1}^{B} |y^i - \hat{y}^i_M| \tag{8}$$

Lastly, we calculate the weighted sum of all training losses to obtain the final optimization objective, which is represented as:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{sem} + \alpha\mathcal{L}_{geo} + \beta\mathcal{L}_{dis} \tag{9}$$

where α and β denote the hyperparameters controlling the impact of the training losses for the missing-modal representations in the feature space.

4 Experiment Setting

4.1 Datasets and Metrics

The experiments are conducted on two benchmark datasets in MSA research: **CMU-MOSI** (Zadeh et al., 2016) contains 2,199 monologue utterances sliced from 93 YouTube movie opinion videos spanning 89 reviewers. We utilize 1,284 utterances for training, 229 utterances for validation, and 686 utterances for testing. **CMU-MOSEI** (Zadeh et al., 2018b) expands the multimodal data into 20k video clips from 3,228 videos in 250 diverse topics collected by 1,000 distinct YouTube speakers. We utilize 16,326 utterances for training, 1,871 utterances for validation, and 4,659 utterances for testing. Both of the datasets are annotated for the sentiment on a Likert scale ranging from -3 to +3, where the polarity indicates positive/negative and the absolute value denotes the relative strength of expressed sentiment.

For the evaluation metrics, we report seven-class classification accuracy (Acc7) for sentiment classification in $[-3, +3]$, as well as binary classification accuracy (Acc2) and weighted F1-score (F1) in two measurement settings as non-negative&negative (non-exclude 0) (Zadeh et al., 2017) / positive&negative (exclude 0) (Tsai et al., 2019a). Moreover, we calculate mean absolute error (MAE) and Pearson correlation (Corr) for the regression difference and correlation between the prediction labels and ground truth.

4.2 Baselines

The MSA baselines are broadly categorized as:

- Simply early and late fusion models: **EF-LSTM** (Williams et al., 2018b), **LF-DNN** (Williams et al., 2018a);
- Tensor-based fusion models: **TFN** (Zadeh et al., 2017), **LMF** (Liu et al., 2018);
- Graph-based fusion model: **Graph-MFN** (Zadeh et al., 2018b);
- Generative and translation-based model: **MFM** (Tsai et al., 2019b), **MCTN** (Pham et al., 2019), **CTFN** (Tang et al., 2021);
- Explicitly intra- and inter-modal dynamics manipulation models: **MFN** (Zadeh et al., 2018a), **MISA** (Hazarika et al., 2020);

Models	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
EF-LSTM	34.5	77.8/79.0	77.7/78.9	0.952	0.651	49.3	80.1/80.3	80.3/81.0	0.603	0.682
LF-DNN	33.6	78.0/79.3	77.9/79.3	0.978	0.658	52.1	78.6/82.3	79.0/82.2	0.561	0.723
TFN	33.7	78.3/80.2	78.2/80.1	0.925	0.662	52.2	81.0/82.6	81.1/82.3	0.570	0.716
LMF	32.7	77.5/80.1	77.3/80.0	0.931	0.670	52.0	81.3/83.7	81.6/83.8	0.568	0.727
MFN	34.2	77.9/80.0	77.8/80.0	0.951	0.665	51.1	81.8/84.0	81.9/83.9	0.575	0.720
Graph-MFN	34.4	77.9/80.2	77.8/80.1	0.939	0.656	51.9	81.9/84.0	82.1/83.8	0.569	0.725
MFM	33.3	77.7/80.0	77.7/80.1	0.948	0.664	50.8	80.3/83.4	80.7/83.4	0.580	0.722
MCTN	33.7	78.7/80.0	78.8/80.1	0.960	0.686	52.0	80.4/83.7	80.9/83.7	0.570	0.728
CTFN	29.9	78.9/80.4	78.7/80.3	0.964	0.683	50.0	80.6/83.2	81.0/83.0	0.582	0.720
MulT	35.0	79.0/80.5	79.0/80.5	0.918	0.685	52.1	81.3/84.0	81.6/83.9	0.564	0.732
MISA	43.5	81.8/83.5	81.7/83.5	0.752	0.784	52.2	81.6/84.3	82.0/84.3	0.550	0.758
MAG-BERT	45.1	82.4/84.6	82.2/84.6	0.730	0.789	52.8	81.9/85.1	82.3/85.1	0.558	0.761
Self-MM	45.8	82.7/84.9	82.6/84.8	0.731	0.785	53.0	82.6/85.2	82.8/85.2	0.540	0.763
MMIM	45.0	83.0/85.1	82.9/85.0	0.738	0.781	53.1	81.9/85.1	82.3/85.0	0.547	0.752
MISA$^{\tau}$	41.3	80.6/82.4	80.6/82.4	0.795	0.764	52.6	81.3/84.8	81.7/84.7	0.545	0.761
MAG-BERT$^{\tau}$	46.0	82.5/84.4	82.4/84.4	0.734	0.790	53.5	81.8/84.8	82.1/84.7	0.542	0.758
Self-MM$^{\tau}$	46.1	82.4/84.2	82.4/84.1	0.727	0.791	53.7	79.6/84.0	80.2/84.0	0.535	0.763
MMIM$^{\tau}$	44.6	83.1/84.3	83.1/84.4	0.753	0.771	53.2	80.4/84.0	80.9/83.9	0.547	0.755
MissModal	**47.2**	**84.1/86.1**	**84.0/86.0**	**0.698**	**0.801**	**53.9**	**83.4/85.9**	**83.6/85.8**	**0.533**	**0.769**

Table 1: Performance comparison between MissModal and baselines with complete modalities on the testing sets of MOSI and MOSEI. $^{\tau}$ denotes that the baselines are reproduced under robust training (Hazarika et al., 2022).

- Transformer-based fusion models: **MulT** (Tsai et al., 2019a), **MAG-BERT** (Rahman et al., 2020);
- Label-guidance: **Self-MM** (Yu et al., 2021);
- Mutual information maximization model: **MMIM** (Han et al., 2021).

We reproduce the baselines with hyperparameter grid searches for the best results. Additionally, we run the state-of-the-art models under robust training with 15% masking and 15% noisy language data following Hazarika et al. (2022) in the circumstances of complete modalities and missing textual modality for a fair comparison.

4.3 Implementation Details

Following the settings of baselines, we adopt the pre-trained BERT-base-uncased model to encode textual input and obtain raw textual features with 768-dimensional hidden states for each token. Besides, we utilize the CMU-Multimodal SDK to pre-process audio and vision data which applies COVAREP (Degottex et al., 2014) and Facet[1] to extract raw acoustic and visual features.

We conduct the experiments on a single GTX 1080Ti GPU with CUDA 10.2. For the hyperparameters, following Gkoumas et al. (2021), we perform fifty-times random grid search to find the best hyperparameters setting including α and β in $\{0.3, 0.5, 0.7\}$, and τ in $\{0.5, 0.7, 0.9\}$. The batch sizes for MOSI and MOSEI are set as 32. For optimization, we adopt AdamW (Loshchilov and Hutter, 2019) as the optimizer with the learning rate 5e-5 for the parameters of BERT on both datasets, and 5e-4 on MOSI and 1e-3 on MOSEI for the other parameters.

For both complete and missing modalities settings, we run experiments five times and report the average performance as the final results. In the experiments with missing modalities, we remain the completeness of modalities of the training sets of both datasets to fine-tune the model, and then freeze the model for the validation and testing sets with different missing rates for diverse modalities

[1] iMotions 2017, `https://imotions.com/`.

Models	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
Missing 30% textual modality										
MISA$^{\tau}$	31.8	72.7/73.8	72.2/73.5	1.003	0.620	48.3	77.2/76.6	76.3/75.1	0.653	0.616
MAG-BERT$^{\tau}$	36.3	70.6/71.0	70.0/70.5	0.994	0.624	47.8	75.3/77.2	75.5/76.6	0.650	0.623
Self-MM$^{\tau}$	35.8	74.1/76.1	73.0/75.3	0.929	**0.669**	**49.1**	77.2/77.6	76.8/76.5	**0.627**	0.634
MMIM$^{\tau}$	34.9	73.3/74.4	73.0/74.2	0.968	0.627	49.0	76.9/77.0	77.0/76.4	0.640	0.628
MissModal	**38.7**	**74.2/76.3**	**73.2/75.5**	**0.907**	0.664	**49.1**	**77.9/78.1**	**77.4/77.0**	0.634	**0.635**
Missing 50% textual modality										
MISA$^{\tau}$	27.3	66.4/67.1	64.7/65.6	1.134	0.525	46.5	75.0/73.2	73.3/70.4	0.696	0.535
MAG-BERT$^{\tau}$	29.5	65.6/66.3	63.7/64.5	1.128	0.526	46.1	74.6/**73.6**	73.4/71.0	0.700	0.533
Self-MM$^{\tau}$	31.8	67.7/69.3	**66.8/67.8**	1.053	0.567	47.2	74.2/72.5	72.7/70.2	**0.691**	**0.539**
MMIM$^{\tau}$	28.8	66.7/67.2	64.4/65.1	1.142	0.530	46.7	75.0/72.7	73.0/70.9	0.697	0.529
MissModal	**32.3**	**67.8/69.6**	65.4/67.4	**1.039**	**0.580**	**47.3**	**75.1/73.6**	**73.5/71.1**	0.692	**0.539**
Missing 70% textual modality										
MISA$^{\tau}$	22.9	61.3/61.7	**59.1**/59.6	1.248	0.389	44.2	72.8/69.1	69.1/62.2	0.763	**0.410**
MAG-BERT$^{\tau}$	22.6	62.1/64.0	55.8/57.9	1.253	0.409	43.8	73.0/68.5	68.4/62.6	0.768	0.392
Self-MM$^{\tau}$	25.5	61.4/62.9	57.7/59.4	1.194	0.442	43.5	72.3/67.1	68.5/62.1	**0.755**	0.404
MMIM$^{\tau}$	23.4	60.4/60.9	55.2/55.9	1.283	0.403	43.9	72.1/68.6	69.0/63.5	0.756	0.405
MissModal	**27.1**	**64.4/66.7**	**59.1/61.8**	**1.164**	**0.450**	**44.6**	**73.9/69.3**	**69.4/63.6**	**0.755**	0.407
Missing 90% textual modality										
MISA$^{\tau}$	19.8	56.9/56.7	46.4/47.1	1.329	0.267	42.2	64.4/60.6	56.3/51.0	0.830	**0.265**
MAG-BERT$^{\tau}$	18.5	55.3/54.8	45.5/46.4	1.416	0.216	42.2	71.6/63.4	62.8/53.9	0.820	0.247
Self-MM$^{\tau}$	18.3	53.6/53.5	44.3/44.4	1.374	0.254	40.7	71.0/62.8	62.8/52.2	0.822	0.250
MMIM$^{\tau}$	18.7	54.3/54.5	44.2/44.6	1.392	0.246	41.8	69.3/64.1	62.7/53.6	0.818	0.239
MissModal	**21.3**	**57.9/60.1**	**48.3/50.9**	**1.316**	**0.271**	**42.4**	**71.9/64.8**	**62.9/54.0**	**0.813**	0.232

Table 2: Performance comparison between MissModal and baselines with missing textual modality in 30%, 50%, 70% and 90% missing rates. $^{\tau}$ denotes the baselines are reproduced under robust training (Hazarika et al., 2022).

to evaluate both flexibility and efficiency of the proposed approach.

5 Experiment Results

5.1 Experiments with Complete Modalities

As shown in Table 1, we compare the performance of MissModal with the state-of-the-art MSA methods with complete modalities in training and testing. The outstanding results on all metrics demonstrate the effectiveness of the proposed architecture of modality-specific and cross-modal representation learning on both MOSI and MOSEI.

Moreover, most previous MSA models demand the presence of fully modalities, which cannot be directly employed when missing modalities in the input data. To address this issue, we adopt robust training (Hazarika et al., 2022) strategy for the state-of-the-art MSA models. However, regardless of the circumstances of missing modalities, we observe that robust training decreases the performance on most metrics when testing with complete modalities due to the introduction of masking or noisy input.

Differently, the superior experiment results of MissModal are achieved under the constraints with the representation of missing modalities, which indicates that the introduction of the missing modality mechanism does not impact the testing performance with complete modalities.

5.2 Experiments when Missing Modalities

To show the benefits from the proposed constraints addressing missing modality issues, we remove modalities by replacing the modality input to zero

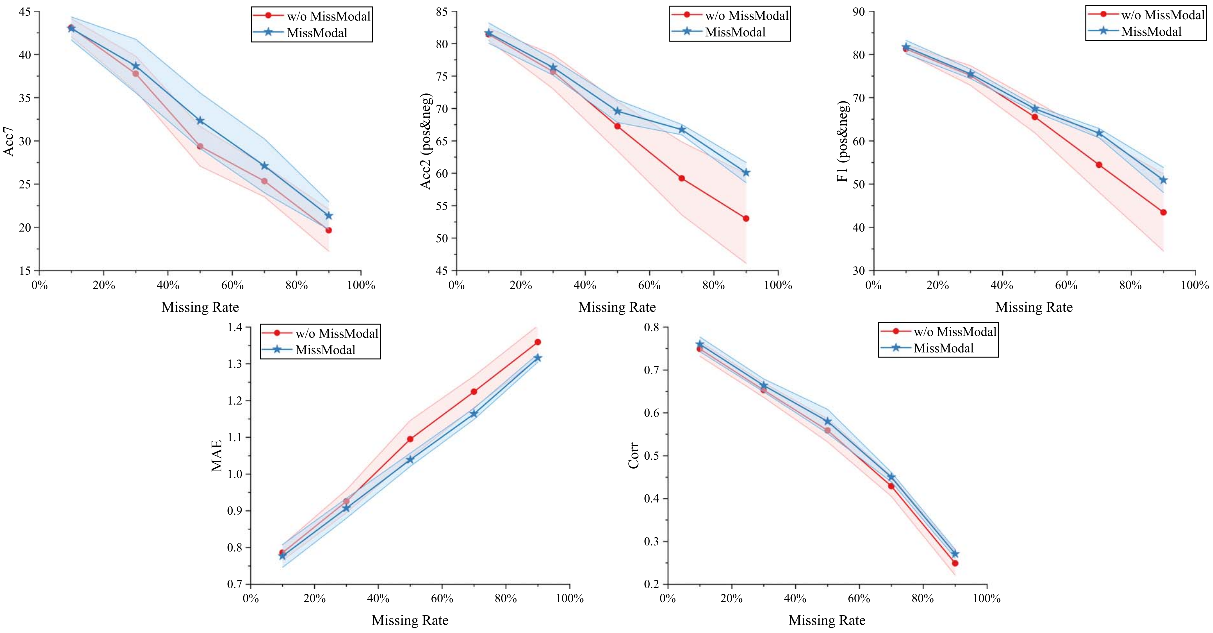


Figure 3: Performance improvement of MissModal in 10%–90% missing rates of textual modality on MOSI.

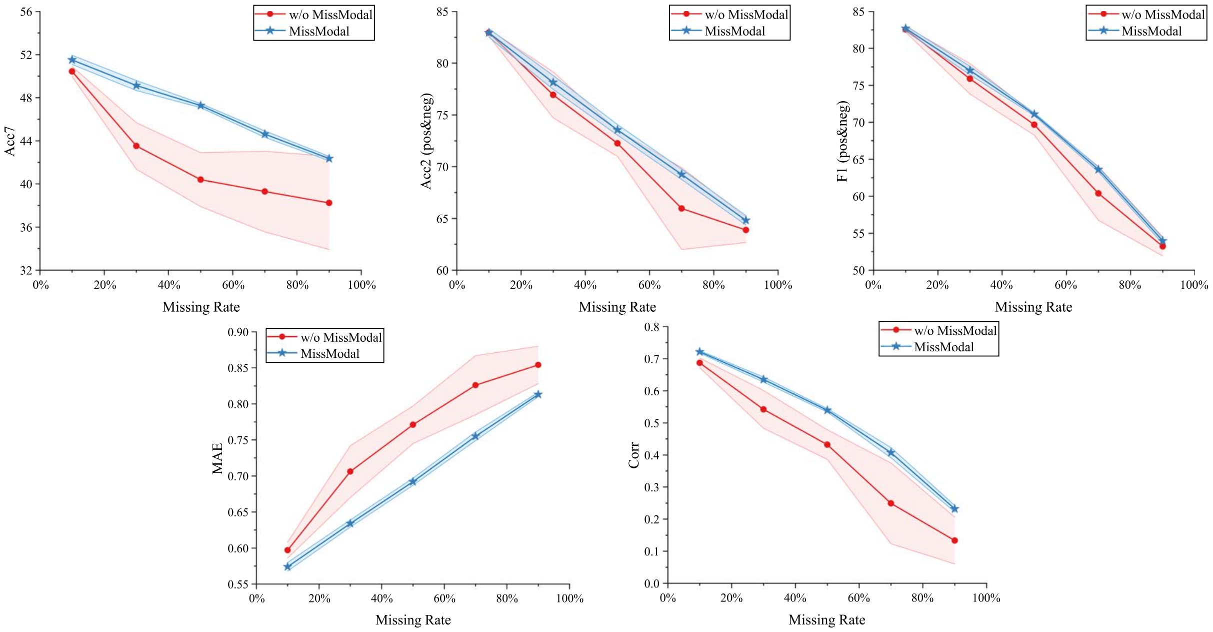


Figure 4: Performance improvement of MissModal in 10%–90% missing rates of textual modality on MOSEI.

vector in both validation and testing sets. Notably, unlike Hazarika et al. (2022) training and testing on language as the specific missing modality, we evaluate MissModal in various scenarios of missing textual modality, missing acoustic or visual modality, and missing random modalities.

5.2.1 Missing Textual Modality

The textual modality is viewed as the dominant modality in MSA task (Hazarika et al., 2020; Wu et al., 2021; Lin and Hu, 2023) due to the large-scale pre-trained language model and the nature of abundant semantic information instrumental in sentiment understanding. We firstly compare MissModal with the state-of-the-art methods under robust training (Hazarika et al., 2022) with missing textual modalities in various missing rates. As shown in Table 2, MissModal achieves superior performance than the state-of-the-art methods under robust training on most metrics, especially in the circumstances of severely missing modalities. We assume that the

Setting	Modality	Missing Rate	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
w/o MissModal	Acoustic	50%	45.9	82.4/84.3	82.3/84.3	0.728	0.782
		90%	45.6	81.8/83.5	81.8/83.6	0.724	0.791
	Visual	50%	45.0	82.1/83.4	82.1/83.5	0.731	0.779
		90%	44.5	81.5/83.1	81.5/83.2	0.737	0.781
MissModal	Acoustic	50%	46.7	**83.2/85.1**	**83.2/85.1**	0.718	**0.800**
		90%	46.1	82.8/84.6	82.8/84.6	**0.709**	0.794
	Visual	50%	**47.4**	**83.2**/84.6	**83.2**/84.6	0.714	0.787
		90%	45.8	82.8/84.4	82.7/84.4	0.723	0.785

Table 3: Performance improvement of MissModal in 50% and 90% missing rates of acoustic and visual modality on the testing set of MOSI dataset.

Setting	Modality	Missing Rate	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
w/o MissModal	Acoustic	50%	53.1	80.2/84.8	80.8/84.8	0.539	0.765
		90%	52.9	79.1/84.0	79.9/84.1	0.541	0.764
	Visual	50%	53.2	80.2/84.8	80.9/84.9	0.539	0.765
		90%	52.7	78.3/83.9	79.2/84.1	0.550	0.761
MissModal	Acoustic	50%	**53.7**	81.9/85.2	82.4/85.1	0.540	**0.768**
		90%	53.2	81.6/84.9	82.1/84.9	0.540	0.766
	Visual	50%	53.2	**82.4/86.1**	**82.9/86.0**	**0.536**	**0.768**
		90%	53.1	81.9/85.6	82.4/85.6	0.539	0.767

Table 4: Performance improvement of MissModal in 50% and 90% missing rates of acoustic and visual modality on the testing set of MOSEI dataset.

fixed settings of missing and noisy rates (15%) of robust training limit its applications on higher missing rates of textual modality. On the contrary, MissModal concentrates on improving the robustness of missing-modal representations, whose performance does not depend on the fixed setting of missing rates.

To further show the effectiveness of MissModal in flexibility and efficiency, we run the model with and without MissModal in different missing rates of textual modality on the testing sets of MOSI and MOSEI datasets as shown in Figures 3 and 4. We observe that missing textual modalities from 10%–90% rates brings more significant drops of average performance to the model without MissModal than the one with MissModal. Besides, the variance of performance without MissModal on all metrics grows rapidly as the increasing of missing rates, which does not happen in the experiment results of the model with MissModal. Moreover, missing textual modality leads to polarization of the predicted sentiment, which is due to the less attention of acoustic and visual modalities to the fine-grained sentiment. Therefore, MissModal helps the model learn more distinguishable missing-modal representations, greatly improving the accuracy of sentiment inference, especially in the case of severely missing modality.

5.2.2 Missing Acoustic or Visual Modality

As the inferior modalities in MSA, acoustic and visual modalities play auxiliary and complementary roles in the prediction of sentiment, leading to less impact on the performance when removing these two modalities at 50% and 90% missing rates on MOSI and MOSEI, as shown in Tables 3–4. Nevertheless, missing each of them bring sub-optimal solution for the MSA model. With missing any modality in any missing rates, the performance of the model with MissModal surpasses the one without MissModal on all metrics, demonstrating the superiority of the proposed approach.

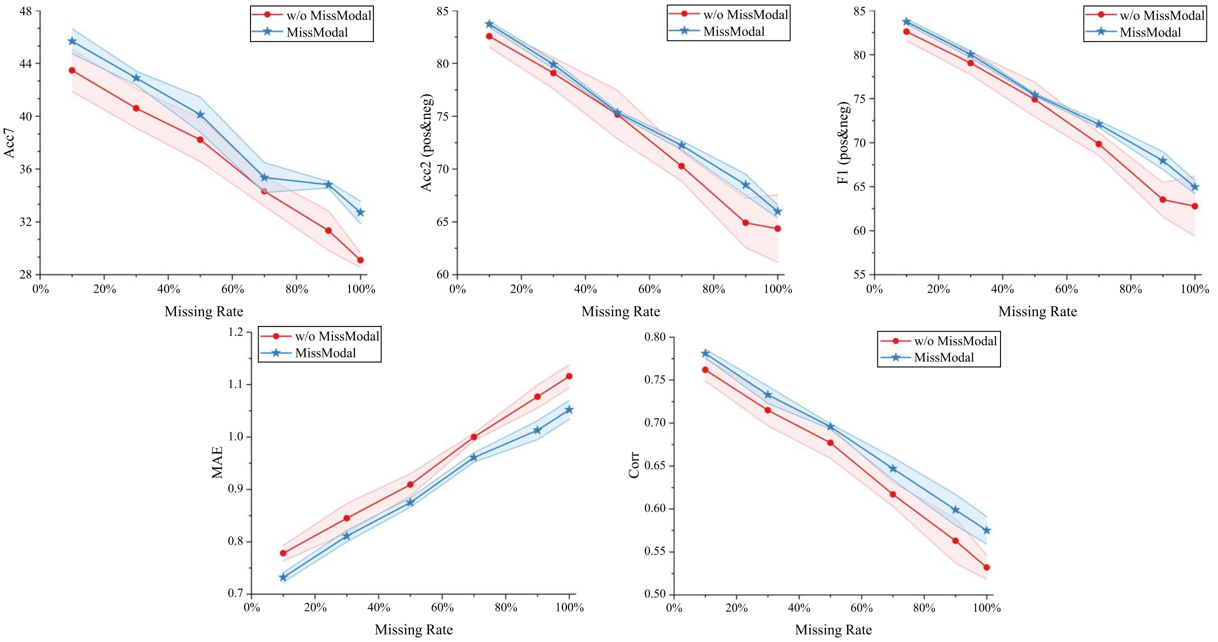


Figure 5: Performance improvement of MissModal in 10%–100% missing rates of random modalities on MOSI.

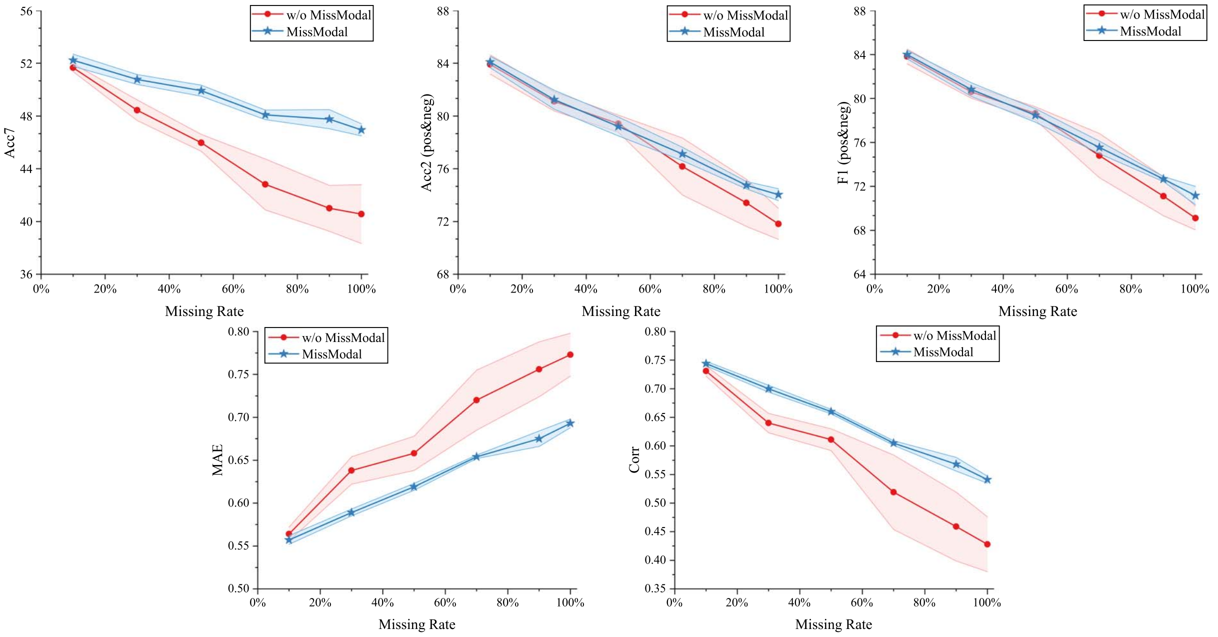


Figure 6: Performance improvement of MissModal in 10%–100% missing rates of random modalities on MOSEI.

5.2.3 Randomly Missing Modalities

To demonstrate wider applications of MissModal in addressing the issues of missing modalities, we remove the modalities in the strategy of random distribution sampling and run MissModal with inputs of the remaining modalities as the settings of $\{(T), (A), (V), (T, A), (T, V), (A, V)\}$. This experiment setting is consistent with the scene when adopting MSA model in the real world where the presence of modalities is unknown.

As shown in Figures 5 and 6, the modalities are randomly removed in various missing rates ranging from 10%–100% on the testing sets of MOSI and MOSEI datasets, where 100% missing rate means that each testing utterance is incomplete and misses modalities randomly. The model with MissModal has a higher average performance and lower variance than the one without MissModal, indicating MissModal remains the upper bound of sentiment prediction performance in the scenarios of missing modalities.

	$\mathcal{L}_{sem}$	$\mathcal{L}_{geo}$	$\mathcal{L}_{dis}$	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
Supervised	✓	✓	✓	**47.0**	**76.0/73.9**	**74.0/71.2**	**0.693**	**0.541**
	✓	✓		46.6	75.3/73.1	73.7/69.8	0.701	0.523
	✓		✓	46.5	74.8/72.0	73.5/69.2	0.714	0.513
	✓			44.4	73.9/72.1	72.3/70.7	0.721	0.503
Unsupervised		✓	✓	34.4	71.8/65.4	67.3/60.0	1.100	0.137
		✓		33.0	66.7/61.9	64.1/58.0	1.048	0.149
			✓	33.7	71.7/64.6	65.9/57.8	1.126	0.148

Table 5: Ablation study of the proposed losses in MissModal with 100% missing rate of random modalities on the testing set of MOSEI dataset, which is divided into supervised and unsupervised circumstances for the learning of missing-modal representations according to the existence of $\mathcal{L}_{sem}$.

Furthermore, it is observed that MissModal has greater improvement in performance and stability on MOSEI than on MOSI, no matter in the settings of missing textual modality or random modalities. We assume that on MOSI, the model tends to overfit the data due to the small scale of the dataset, while on MOSEI, the larger data scale helps reveal more significant improvement on the generalization performance of the proposed approach.

In general, MissModal reaches more stable and superior performance in the experiments both on flexibility as the randomness of missing modalities and on efficiency, as severely missing modalities at even 100% randomly missing rate.

6 Further Analysis

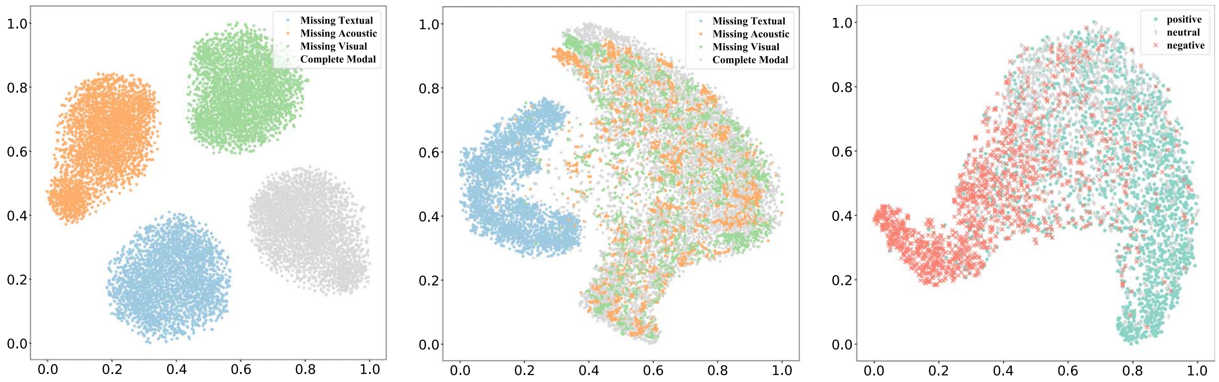
6.1 Ablation Study

We conduct an ablation study on the proposed losses $\mathcal{L}_{sem}$, $\mathcal{L}_{geo}$, and $\mathcal{L}_{dis}$ of MissModal with 100% missing rate of random modalities, as shown in Table 5. Apparently, each loss contributes to the training and encourages the model to reach optimal performance. Besides, with the supervision of the ground truth labels in $\mathcal{L}_{sem}$, MissModal achieves greatly higher performance than the one trained without $\mathcal{L}_{sem}$. Nevertheless, only $\mathcal{L}_{sem}$ guiding the learning at the level of prediction is far not enough for the representations when missing modalities. By fine-tuning at the feature level with $\mathcal{L}_{geo}$ and $\mathcal{L}_{dis}$, the model learns more robust miss-modal representations. Intriguingly, both $\mathcal{L}_{geo}$ and $\mathcal{L}_{dis}$ can enhance the performance of miss-modal representations even in the unsupervised circumstance without the assistance of $\mathcal{L}_{sem}$, which reveals new sight for the field of unsupervised MSA.

Modality	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
only T	52.8	80.9/84.7	81.4/84.6	0.545	0.761
only A	41.4	71.0/62.9	59.0/48.5	0.839	0.038
only V	41.4	70.2/62.3	59.3/49.2	0.840	0.018
Complete	**53.9**	**83.4/85.9**	**83.6/85.8**	**0.533**	**0.769**

Table 6: Ablation study of various modalities in MissModal on the testing set of MOSEI dataset. T, A, V denote textual, acoustic, and visual modality.

Additionally, we evaluate the performance of MissModal with only one specific modality when totally missing information from other modalities. As shown in Table 6, the experiment illustrates that textual modality is the dominant modality while acoustic and visual modalities serve as the inferior modalities in MSA task, concluding consistent with the former results and previous research (Gkoumas et al., 2021). However, only textual modality may trap the model into the subjective and biased emotion problems (Zadeh et al., 2017; Wang et al., 2019), degrading the performance compared with the multimodal case. Thus, the introduction of acoustic and visual modalities is necessary to further boost the accuracy of sentiment inference for the MSA task. Each modality of the utterance provides unique and complementary properties, which are extracted as modality-specific and -shared features for the final sentiment prediction. The demand for various modalities indicates the necessity of improving the robustness of MSA models when missing modalities.



(a) Before MissModal (b) After MissModal (c) Over Sentiment Classes

Figure 7: Visualization of (a)(b) multimodal representation with missing and complete modalities in the embedding space on the training set of MOSEI and (c) multimodal representation with complete modalities over different sentiment classes in the embedding space on the testing set of MOSEI.

#	Multimodal Input Example (T+A+V)	Ground Truth	MissModal	w/o MissModal
1	~~"More power to him, I don't blame him"~~ + Calm and firm tone + Shaking head and relieved facial expression	0.0	0.083	1.165
2	"The Treu Group is having an amazing success in 'the farms' and we would love to help you, too; if you're thinking of selling"+ ~~Positive and emphasis tone~~ + Smiling face and active expression	1.667	1.659	2.048
3	"Well someone says something and the next moment you start rolling your eyes with anger" + Annoyed and disappointed tone + ~~Twisted, frustrated and frown face~~	-0.667	-0.612	-0.358
4	"(umm) Awful awful acting" + ~~Upset and emphasis tone~~ + ~~Compromised and resigned facial expression~~	-2.333	-2.372	-2.696
5	"So basically I had a friend, I'm not going to mention any real names here, but for the sake of clarification let's say her name was Madison" + ~~Active and passionate tone~~ + ~~Focused and serious face~~	0.333	0.276	0.069

Table 7: Examples from the testing set of CMU-MOSEI dataset. The missing modalities input is highlighted with red ~~strikethrough~~ and the ground truth sentiment labels are between strongly negative (-3) and strongly positive (+3). For each example, we show the Ground Truth and output predictions of models with and without MissModal.

6.2 Representation Visualization

As shown in Figure 7(a)–7(b), we utilize the t-SNE algorithm (Van der Maaten and Hinton, 2008) to provide visualization in the embedding space for the learning processes of missing and complete representations. Before training, significant modality gaps exist among the missing-modal and complete-modal representations. Through the guidance of three proposed constraints, MissModal successfully aligns the distributions of the representations with missing acoustic or visual modalities and the ones with complete modalities, leading to superior results in the experiments with missing acoustic and visual modalities in Tables 3 and 4. Nevertheless, we observe that the absence of semantic information makes it challenging to optimize and align the multimodal representations lacking the textual modality, highlighting the dominant role of textual modality as indicated by the results in Table 6. Despite the remaining gaps in the embedding space, the distribution shape of representations without textual modality is similar to others in Figure 7(b), illustrating the effectiveness of MissModal even in the circumstance of missing the dominant modality.

Furthermore, we visualize the representations over different sentiment classes with complete modalities in the downstream testing in Figure 7(c) to demonstrate the superiority of MissModal in the downstream inference. The learned multimodal representations are divided into distinguishable clusters according to positive, neutral, and negative sentiment. Besides, the representations inside the same sentiment class are compact and become more and more compact with the increasing sentiment intensity. This reveals the relation between multimodal representations and sentiment labels, implicitly indicating the productive collaboration among $\mathcal{L}_{geo}$, $\mathcal{L}_{dis}$ in the feature space, and $\mathcal{L}_{sem}$ at the level of prediction.

6.3 Qualitative Analysis

To further validate the contribution of the proposed approach, we present some examples where MissModal achieves superior performance compared with the model without MissModal when missing modalities in the multimodal input data in Table 7. The examples show various circumstances of missing modalities to demonstrate the effectiveness of the three proposed constraints.

Examples 1 to 3 contain multimodal input with missing only one modality, where the missing modality provides additional information to the final sentiment prediction. Without these complementary information, the model without MissModal tends to over-amplificate or over-reduce the magnitude of emotion contained in the utterances. Diversely, MissModal aligns the missing-modal representations with the complete-modal ones in the training, which implicitly transfer the knowledge of the missing modality to the remaining ones in the guidance of sentiment labels. Thus, the sentiment prediction of MissModal is closer to the annotated ground truth label in these cases, leading to higher performance on Acc7, MAE, and Corr as shown in Figure 6.

Examples 4 and 5 show cases without both acoustic and vision modalities, illustrating that these two inferior modalities present auxiliary roles in sentiment inference. Especially in Example 5, the text in the utterance can be potrayed as mostly neutral, which results in a prediction score close to 0 for the model without MissModal. While due to the latent information conveyed by the active tone and focused facial expression, MissModal deflects the polarity of sentiment to a bit positive, similar as the given ground truth label.

Model	Increased Parameters
MFM	5,996,541
MCTN	3,910,658
CTFN	2,852,801
MISA$^{\tau}$	1,435,105
MAG-BERT$^{\tau}$	1,352,449
Self-MM$^{\tau}$	165,668
MMIM$^{\tau}$	338,889
MissModal	**340,039**

Table 8: Comparison of model complexity of MissModal and the MSA baselines. To highlight the demands for parameters, the reported number of parameters are the increased number of extra parameters after removing the pre-trained language model BERT.

6.4 Model Complexity

As shown in Table 8, we compare the model complexity of various models by reporting the increased number of parameters on CMU-MOSEI.

Firstly, the generative models such as MFM, MCTN, and CTFN require massive parameters as mentioned above, strengthening the motivation of adopting classification-based methods in computationally limited scenarios. Differently, by simplifying the multimodal fusion networks to significantly reduce the computation complexity, MissModal requires parameters less than or comparable with the state-of-the-art baselines. The extra increased parameters for MissModal are brought mostly by multiple fusion networks for various circumstances of missing modalities.

Besides, the proposed constraints in MissModal demand no extra training parameters when addressing the issues of missing modalities. In general, MissModal achieves better trade-off between model complexity and performance with both complete and missing modalities.

7 Limitations

The limitations of the proposed approach are listed in the following for future research. First, the model parameters of MissModal depend on the complexity of multimodal fusion networks and the number of modalities, which may bring problems with increasing model complexity in the downstream application of the proposed approach. Then, the improvement of MissModal seems relevant to the scale of datasets, where small datasets may limit the robustness of MissModal. We believe that increasing the scale of datasets can show more effectiveness of the proposed approach in the issues of missing modalities. Lasty, although MissModal aims at handling missing modalities in the stage of inference, the demand for complete modalities in training raises difficulty in collecting multimodal data. Getting rid of the completeness of modalities in training is another interesting research area for us to explore in the future.

8 Conclusion

In this paper, we present a novel classification-based approach named MissModal to enhance the robustness to missing modalities in the downstream application by constructing three constraints, including geometric contrastive loss, distribution distance loss, and sentiment semantic loss to align the representations of missing and complete modalities. Extensive experiments on various settings of missing modalities and missing rates demonstrate the superiority of MissModal in both flexibility and efficiency on two public datasets. The analysis of representation visualization and model complexity further indicates the huge potential and generality of MissModal in other multimodal systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 62076262.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443. https://doi.org/10.1109/TPAMI.2018.2798607, PubMed: 29994351

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994. https://doi.org/10.1162/tacl_a_00408

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. https://doi.org/10.1109/ICASSP.2014.6853739

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. 2021. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197. https://doi.org/10.1016/j.inffus.2020.09.005

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.723

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies*, pages 685–696, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.50

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131. https://doi.org/10.1145/3394171.3413678

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585. https://doi.org/10.1162/tacl_a_00385

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.202

Ronghao Lin and Haifeng Hu. 2023. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–18. https://doi.org/10.1109/TAFFC.2023.3282410

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1209

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18156–18165. https://doi.org/10.1109/CVPR52688.2022.01764

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2302–2310. https://doi.org/10.1609/aaai.v35i3.16330

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 169–176, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2070481.2070509

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899. https://doi.org/10.1609/aaai.v33i01.33016892

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica

Kragic. 2022. Geometric multimodal contrastive representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17782–17800. PMLR.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*. `https://doi.org/10.1109/TAFFC.2020.3038167`

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.214`, PubMed: 33782629

Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong. 2021. CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5301–5311, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.acl-long.412`

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P19-1656`

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. In *International Conference on Learning Representations*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Miguel Vasco, Hang Yin, Francisco S. Melo, and Ana Paiva. 2022. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255. `https://doi.org/10.1016/j.neunet.2021.11.019`, PubMed: 34906760

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223. `https://doi.org/10.1609/aaai.v33i01.33017216`

Jennifer Williams, Ramona Comanescu, Oana Radu, and Leimin Tian. 2018a. DNN multimodal fusion techniques for predicting video sentiment. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 64–72, Melbourne, Australia. Association for Computational Linguistics. `https://doi.org/10.18653/v1/W18-3309`

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018b. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3302

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.417

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797. https://doi.org/10.1609/aaai.v35i12.17289

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1115

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. https://doi.org/10.1609/aaai.v32i1.12021

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1208

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