Multimodal Invariant Sentiment Representation Learning

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, Ning An
School of Computer Science and Information Engineering, Hefei University of Technology
zhuaoqiang@mail.hfut.edu.cn, {jsjxhumin, xh_wang, jiaoyun, ymtang}@hfut.edu.cn, ning.g.an@acm.org

Abstract

Multimodal Sentiment Analysis (MSA) integrates diverse modalities to overcome the limitations of unimodal data. However, existing MSA datasets commonly have significant sentiment distribution imbalances and cross-modal sentiment conflicts, which hinder performance improvement. This paper shows that distributional discrepancies and sentiment conflicts can be incorporated into the model training to learn stable multimodal invariant sentiment representation. To this end, we propose a Multimodal Invariant Sentiment Representation Learning (MISR) method. Specifically, we first learn a stable and consistent multimodal joint representation in the latent space of Gaussian distribution based on distributional constraints. Then. under invariance constraint, we further learn multimodal invariant sentiment representations from multiple distributional environments constructed by the joint representation and unimodal data, achieving robust and efficient MSA performance. Extensive experiments demonstrate that MISR significantly enhances MSA performance and achieves new state-of-the-art. The code has been released at https://github.com/aoqzhu/MISR.

1 Introduction

Multimodal Sentiment Analysis (MSA) understands human emotions by fusing information from modalities such as vision, audio, and text (Gandhi et al., 2023). MSA effectively compensates for the limitations of a unimodal data by exploring the relationships between different modalities (Xu et al., 2023), and shows significant advantages in improving the understanding and expression of sentiment.

MSA research has focused on utilizing various strategies to enhance performance. Early methods focused on fusion mechanisms for modality integration (Zadeh et al., 2017; Yang et al., 2022), followed by attention-based approaches that enhanced performance by capturing intra- and inter-modal correlations (Zhu et al., 2024; Yang et al., 2023; Lv et al., 2021a). Recently, decoupled representation learning has emerged as a key direction to minimize interference by separating modality information(Sun et al., 2023; Li et al., 2023). Despite achieving impressive improvements, these methods ignore the significant distribution imbalance and cross-modal sentiment conflicts present in the datasets used for model training, limiting the performance of the models.

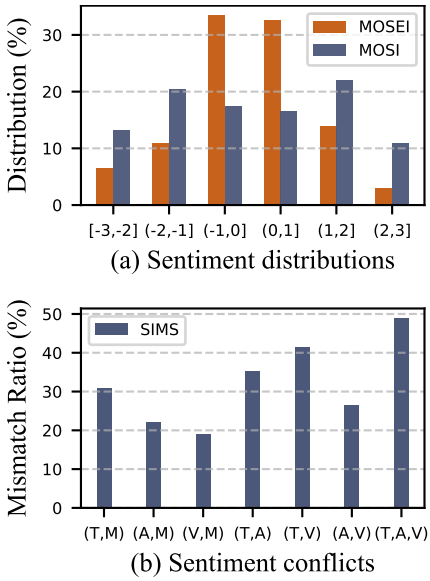

Figure 1: Datasets sample distribution analysis. T, A, V and M denote the text, audio, visual and multimodal sentiment labels of the samples.

As shown in Fig. 1(a), it is evident that in the MOSEI dataset (Bagher Zadeh et al., 2018), sentiment values predominantly fall within the range of [-1,1], accounting for 65.97% of the samples. The imbalance in the sentiment distribution of the data causes the model to bias toward predictions of high-frequency sentiments, while insufficiently learning low-frequency sentiments (Dixon et al., 2018; Qian et al., 2021). Due to the intrinsic uncertainty of the data (Kendall and Gal, 2017), there

Findings of the Association for Computational Linguistics: ACL 2025, pages 14743–14755

are conflicting sentiment across modalities. SIMS (Yu et al., 2020) provides sentiment categories for each modality, and we analyze their inconsistency across modalities. As shown in Fig. 1(b), significant sentiment conflicts exist within unimodal data, with 48.97% inconsistency across the three modalities. This conflict leads to contradictions between modalities, interfering with the fusion process.

Data distribution imbalances are difficult to avoid in real-world training, and countering them is often costly (Tang et al., 2022; Arjovsky et al., 2019). Even when addressed, cross-modal sentiment conflicts persist. Inspired by the Invariant Risk Minimization (IRM) (Tang et al., 2022; Zhou et al., 2023), we observe that while unimodal sentiment knowledge is a primary source for generating multimodal labels, the correlation between unimodal sentiment categories and multimodal labels is unstable (as shown in Fig. 1(b)). We wish to train the MSA model to learn invariant features, stably correlated with multimodal labels (e.g., object shapes in image classification), rather than spurious features (e.g., environments) with unstable correlations (Arjovsky et al., 2019). If the uncertain unimodal sentiment distribution is treated as an unstable spurious feature, how can invariant features directly related to multimodal labels be represented? We propose a reasonable assumption: multimodal sentiment representation is invariant to unimodal sentiment polarity and data distribution. For example, the multimodal sentiment polarity of "positive" remains positive, regardless of changes in unimodal sentiment or data distribution. Based on this, we reformulate the MSA task as learning stable and invariant sentiment representation from imbalanced multimodal data, using invariant constraints to enhance MSA performance.

Based on the above analysis, we propose the Multimodal Invariant Sentiment Representation Learning (MISR). Specifically, we first map data to the latent space of Gaussian distributions to learn stable and consistent joint representation. Then, we treat unimodal data as unstable features relevant to the task and combine them with joint representation to create multiple distribution environments. Finally, based on invariant risk constraints, we learn multimodal invariant sentiment representation across these environments, thereby achieving effective MSA. In the MISR learning process, unfavorable factors in the data distribution are incorporated into optimization objectives for model training. Data distribution imbalance provides an optimization direction for distribution constraints, while sentiment conflicts combined with joint representation construct diverse sentiment distribution environments for invariant learning. MISR learns stable and invariant sentiment representation through joint distribution constraint and invariance constraint, achieving effective and robust MSA performance. The main contributions are as follows:

- We propose the Multimodal Invariant Sentiment Representation Learning (MISR) method, which is based on data distribution discrepancies and sentiment conflicts to learning multimodal invariant sentiment representation, enabling robust and effective MSA.
- We propose a stable multimodal joint representation learning method, which learns stable and semantically consistent joint representation based on a Gaussian distribution-based latent space.
- We propose a multimodal invariant risk minimization theory, extending unimodal invariant feature learning to multimodal invariant representation learning.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) aims to endow machines to understand human emotions through different modalities (Singh et al., 2024). Mainstream MSA methods can be categorized into two types: modality fusion and unified representation learning. Modality fusion methods focus on acquiring multimodal features through complex fusion strategies or mechanisms(Liu et al., 2018; Rahman et al., 2020; Tsai et al., 2019; Wang et al., 2024a; Huang et al., 2023). LMF (Liu et al., 2018), as a typical work, used a low-rank tensor for multimodal fusion and optimized computational efficiency with modality-specific factors. In contrast, unified representation learning focuses on analyzing intra-modal and inter-modal contextual relationships to learn unified multimodal representations (Yu et al., 2023; Zhu et al., 2024; Hazarika et al., 2020; Qian et al., 2023; Zhang et al., 2023). For example, ConKI (Yu et al., 2023) optimized the learning of joint multimodal representations through contrastive knowledge injection. Despite achieving impressive improvements, these methods overlook distribution imbalance and cross-modal

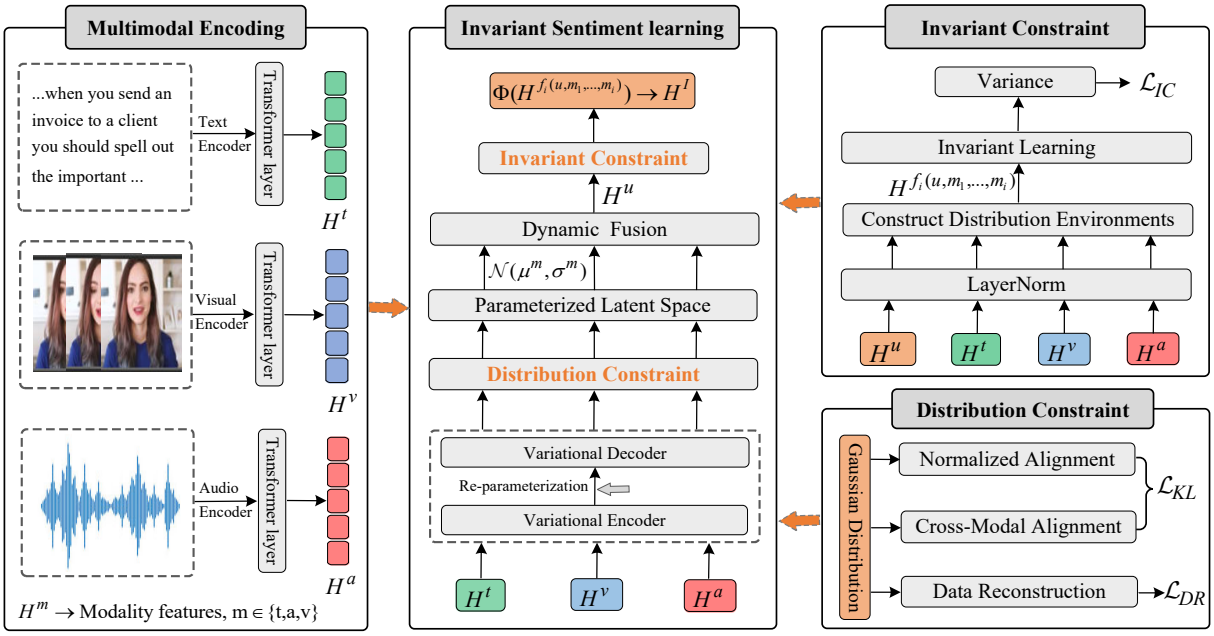

Figure 2: The framework of the Multimodal Invariant Sentiment Representation Learning (MISR).

sentiment conflicts in the training datasets, limiting model performance.

2.2 Invariant Risk Minimization

Invariant Risk Minimization (IRM) is a training method that improves model generalization by learning domain-invariant features (Rosenfeld et al., 2021). The key idea is to constrain the learning process, enabling the model to learn invariant features across domains (Chen et al., 2023; Wang et al., 2024b; Lai and Wang, 2024; Wu et al., 2024). For example, Causal-Debias (Zhou et al., 2023) leveraged specific downstream tasks to identify factors related to bias and labels, mitigating bias from the perspective of causal invariance. Under IRM constraints, the model focuses on task-relevant factors, improving generalization across different distributions. IFL (Tang et al., 2022) learned invariant features across different image data distributions to address class and attribute imbalances. Inspired by this, we propose that invariant feature learning in unimodal can be further extended to invariant representation learning in multimodal settings. Therefore, we propose the Multimodal Invariant Sentiment Representation Learning (MISR) method.

3 Methodology

3.1 Overall Framework

Fig. 2 shows the main modules and workflow of the proposed MISR. MISR consists of four core modules: Multimodal Encoding (ME), Invariant Sentiment Learning (ISL), Invariant Constraint (IC), and Distribution Constraint (DC). Specifically, the ME layer passes the extracted modality features through two layers of Transformer embedding to standardize the different feature dimensions. The ISL module combines Gaussian distribution and Invariant Risk Minimization to learn stable and invariant sentiment representation. The DC motivates the stability and consistency of the multimodal joint sentiment representation learned from the latent space of the Gaussian distribution. Meanwhile, the IC learns invariant sentiment representation from the constructed diverse distribution environments by combining the joint representation. Finally, the learned invariant sentiment representation is used as multimodal features for downstream multimodal sentiment or emotion analysis tasks.

3.2 Multimodal Encoding

We conduct experimental on four multimodal benchmark datasets: MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), SIMS (Yu et al., 2020), and CHERMA (Sun et al., 2023). Following previous works (Yu et al., 2023; Yang et al., 2023; Zhang et al., 2023), we use BERT (Devlin et al., 2019) for text, librosa (McFee et al., 2015) for audio, and OpenFace (Baltrusaitis et al., 2018) for video feature extraction. To ensure consistent feature dimensions across modalities, we standardize them by applying an embedding layer with two Transformer layers to each modality. Given a mul-

timodal input, we denote the multimodal feature sequence as $H^m \in \mathbb{R}^{L_m \times d_m}$. Here, $m \in \{t, a, v\}$ denotes the modality type (i.e., text, audio, and visual), L_m denotes the sequence length of the modality, and d_m denotes the dimensionality of the modality vector.

3.3 Invariant Sentiment Learning

The imbalance in sentiment distribution of the MSA dataset causes the model to favor high-frequency sentiments while insufficiently low-frequency ones. Meanwhile, cross-modal sentiment conflicts further interfere with the fusion process, limiting overall model performance. Inspired by success of IRM in single-modality invariant feature learning (Tang et al., 2022; Zhou et al., 2023), we extend IRM to multimodal invariant sentiment representation learning to address the impact of data distribution on MSA. Due to the heterogeneity of multimodal data, it lacks the obvious invariant features (e.g., object shapes in image classification) and abundant spurious features (e.g., environmental) typically found in single-modality data. Therefore, we emphasize that MISR learns the joint invariant sentiment representation of multimodal data. As shown in Fig. 2, we achieve this in two steps: first, learning stable and consistent joint representations from multimodal Gaussian distribution; second, under the IRM constraint, learning sentiment invariance across various distribution environments created by unimodal data and joint representations. Each part is detailed below.

As shown in Fig. 2, we adopt the concept of Variational Inference and design a Gaussian-based Variational Encoder (VE) and Variational Decoder (VD) for each modality. The variational encoder consists of a mapping layer with ReLU, two Transformer layers, and two fully-connected layers, and is designed to learn the parametric mean $u(h^m)$ and variance $\sigma(h^m)$ of the Gaussian distribution. Given a unimodal sample data h^m, the variational posterior distribution of different modalities in the Gaussian latent space can be expressed as:

$$q(z^m|h^m) \sim \mathcal{N}[z^m \big| u(h^m), \sigma^2(h^m)I] \,, \quad (1)$$

where $z^m, m \in \{t, v, a\}$ represents the latent variable capturing the core features of modality m, and I is the identity matrix. The mean $u(h^m)$ of the Gaussian distribution represents the stable representation in the latent space, while the variance $\sigma(h^m)$ reflects the uncertainty of the distribution. The discrepancies in unimodal sentiment contribution reflects modality distribution uncertainty, which can be quantified through the uncertainty in the feature space distribution. We quantify the fusion weights ω^m of cross-modal representations based on the modality uncertainty representation $\sigma(h^m)$.

$$\omega^m = \frac{\exp(1/\sigma(h^m))}{\sum_{m \in M} \exp(1/\sigma(h^m))}, \quad (2)$$

where $M = \{t, v, a\}$. Given the stable unimodal representation $u(h^m)$, we apply the quantified uncertainty fusion weights ω^m to the dynamical fusion of the joint multimodal representation.

$$H^u = \sum_{m \in M} \omega^m \cdot u(h^m), \quad (3)$$

where H^u denotes the stable and consistent joint multimodal representation. To optimize the distribution and smooth and stabilize the learning process of H^u, we introduce the standard normal distribution $\mathcal{N}(0, I)$ as a priori constraint. By minimizing the KL divergence to align the posterior distribution $q(z^m|h^m)$ with $\mathcal{N}(0, I)$, promoting the stability of the model in learning H^u.

$$\mathcal{L}_{NA} = \sum_{m \in M} D_{KL}(q(z^m|h^m)|\mathcal{N}|(0, I)) \,, \quad (4)$$

where $\mathcal{L}_{NA}$ is the normalized alignment constraint loss. To align the distributions across modalities and promote consistent representations, we introduce a cross-modal KL divergence alignment constraint, as follows:

$$\mathcal{L}_{CMA} = \sum_{m_1 \neq m_2} D_{KL}(q(z^{m_1}|h^{m_1})|q(z^{m_2}|h^{m_2})), \quad (5)$$

where $m_1, m_2 \in M$, $\mathcal{L}_{CMA}$ denotes the cross-modal alignment constraint loss. To make the latent space sampling process differentiable and enable effective gradient propagation, we employ the reparameterization trick to sample from the Gaussian distribution in the latent space. Specifically, we represent the latent variable z^m as:

$$z^m = u(h^m) + \sigma(h^m) \cdot \varepsilon, \quad (6)$$

where $\varepsilon \sim \mathcal{N}(0, I)$ is the noise sampled from the standard normal distribution. The latent variable z^m serves as the core abstract representation of the input data h^m, reflecting the effectiveness of the latent space parameterization. Therefore, we

design a Variational Decoder (VD) for each modality, consisting of a mapping layer with ReLU, two Transformer layers, and a fully connected layer to reconstruct the input data h^m from z^m.

$$\mathcal{L}_{DR} = \mathbb{E}_{q(z^m|h^m)}[||h^m - \text{VD}(z^m)||^2] \quad (7)$$

Thus, the overall distribution constraint $\mathcal{L}_{DR}$ of the Gaussian distribution can be expressed as:

$$\mathcal{L}_{DC} = \mathcal{L}_{NA} + \mathcal{L}_{CMA} + \mathcal{L}_{DR} \quad (8)$$

The distributional discrepancies in the data provide optimization objective for the $\mathcal{L}_{DC}$, promoting the consistency and stability of the multimodal joint representation H^u. Inspired by IRM (Tang et al., 2022; Zhou et al., 2023), we argue that a consistent and stable multimodal joint representation H^u can further learn invariant sentiment representation under the IRM constraint. In Section 3.4, we detail the construction and learning process of the multimodal invariant risk minimization.

3.4 Invariant Constraint

In Section 3.3, we obtain a consistent and stable multimodal joint representation H^u. Based on previous analysis, H^u serves as a direct cause of multimodal sentiment label Y. We regard H^u as an invariant representation of multimodal. Furthermore, cross-modal sentiment conflicts in data provide diverse distributional environments for invariant learning. Thus, we combine H^u and unimodal representation $H^m, m \in M$ to construct multiple distributional environments .

$$\varepsilon = \{e_1, e_2, ..., e_E\} \in \sum H^{f_i(u,m_1,...,m_i)}, \quad (9)$$

where $i \in \{0, 1, 2, 3\}$, f_i is fusion function. For example, $H^{f_2(u,m_1,m_2)}$ denotes the joint representation of H^u with H^{m_1} and H^{m_2}, $m_1, m_2 \in \{t, v, a\}$ and $m_1 \neq m_2$. The special $f_0 = u$. In the MSA task, we wish to learn a multimodal sentiment distribution $\Phi(H)$ that depends only on H^u. If $\Phi(H)$ is solely dependent on H^u, meaning its representation does not rely on the environment and is a task-relevant joint-invariant representation H^I, then for any distributional environment e, we have:

$$P^e(Y|\Phi(H)) = P(Y|\Phi(H)) \ , \quad (10)$$

where $H \in H^{f_i(u,m_1,...,m_i)}$. Since $\Phi(H)$ is learned as a joint representation solely dependent on H^u, i.e., the invariant representation H^I. Therefore $P(Y|\Phi(H))$ denotes the optimal prediction for the task label Y, while $P^e(Y|\Phi(H))$ denotes the prediction for Y in the environment e. As $\Phi(H)$ is still learned as a joint representation only relying on H^u in environment e, it holds that $P^e(Y|\Phi(H)) = P(Y|\Phi(H))$ for any environment e.

Under the core idea of IRM, we constrain the model to learn invariant representations across environments with the aim of achieving optimal results in all environments (Lai and Wang, 2024; Wu et al., 2024). According to the IRM optimization algorithm in REx (Krueger et al., 2021), if $\Phi(H)$ represents the invariant representation learned by the model across different environments e, the prediction loss for the sample label Y remains the same across environments. Therefore, the model can be optimized by minimizing the loss variance (Var) across environments. As a result, the multimodal invariant constraint $\mathcal{L}_{IC}$ optimization loss function is expressed as:

$$\mathcal{L}_{IC} = \text{Var}(R^e(\theta, \Phi(\sum H^{f_i(u,m_1,...,m_i)}))), \quad (11)$$

where $R^e(\theta, \Phi)$ denotes the risk under environment e, which in this paper refers to the absolute error loss of the model prediction, and θ denotes the model parameters. In the IRM framework, cross-modal sentiment conflicts are used to construct multiple distributional environments for invariant learning, optimizing the model through invariance constraints and enhancing its robustness.

3.5 Overall Learning Objectives

The learned multimodal invariant sentiment representation H^I is fed into a Multi-Layer Perceptron (MLP) consisting of two fully connected layers with ReLU, which outputs the sentiment prediction. The model is optimized using L1 loss to minimize sentiment prediction error.

$$\mathcal{L}_{SP} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \quad (12)$$

where $\mathcal{L}_{SP}$ denotes the sentiment prediction loss, $\hat{y}_i = \text{MLP}(H_i^I)$ denotes the predicted value of sample i, and y_i denotes the sample label. In particular, the cross-entropy loss is used for classifying seven emotions in the CHERMA (Sun et al., 2023).

In summary, the total optimization objectives of MISR can be expressed as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{SP} + \lambda_2\mathcal{L}_{DC} + \lambda_3\mathcal{L}_{IC}, \quad (13)$$

where λ_1, λ_1, and λ_1 are weight hyperparameters.

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
LFN*	-/80.8	-/80.7	34.9	0.901	0.698	-/82.5	-/82.1	50.2	0.593	0.700
LMF*	-/82.5	-/82.4	33.2	0.917	0.695	-/82.0	-/82.1	48.0	0.677	0.623
MISA*	80.79/82.10	80.77/82.03	-	0.804	0.764	82.59/84.23	82.67/83.97	-	0.548	0.724
CMHFM	79.45/81.10	79.30/81.02	42.13	0.822	0.723	84.31/84.37	84.18/84.01	52.61	0.548	0.747
TMBL	81.78/83.84	82.41/84.29	36.3	0.867	0.762	84.23/85.84	84.87/85.9	52.4	0.545	0.766
MAG-BERT*	82.37/84.43	82.50/84.61	43.62	0.781	0.727	82.51/84.82	82.77/84.71	52.67	0.543	0.755
Self-MM*	82.54/84.77	82.68/84.9	45.79	0.712	0.795	82.68/84.96	82.95/84.93	53.46	0.529	0.767
HyCon*	-/85.2	-/85.1	46.6	0.713	0.790	-/85.4	-/85.6	52.8	0.601	0.776
MMIM*	84.14/86.06	84.00/85.98	46.65	0.700	0.800	82.24/85.97	82.66/85.94	54.24	0.526	0.772
ConKI*	84.37/86.13	84.33/86.13	48.43	0.681	0.816	82.73/86.25	83.08/86.15	54.25	0.529	0.782
ALMT	84.55/86.43	84.57/86.47	49.42	0.683	0.805	84.78/86.79	85.19/86.86	54.28	0.526	0.779
MISR	**86.01/88.11**	**86.10/88.15**	**49.85**	**0.671**	**0.819**	**85.28/87.51**	**85.57/87.55**	**55.05**	**0.513**	**0.789**

Table 1: Performance of models on MOSI and MOSE. "*" indicates the results are from ConKI (Yu et al., 2023). The best results are marked in bold.

4 Experiments

4.1 Datasets

We evaluate the MISR model on two tasks: Multimodal Sentiment Analysis (MSA) and Multimodal Emotion Recognition (MER).

MSA: MSA aims to analyze the sentiments expressed by people in a video. We evaluate on the MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and SIMS (Yu et al., 2020) datasets.

MER: MER focuses on classifying verbal emotions in videos into multiple emotion categories. We evaluate on the CHERMA (Sun et al., 2023) dataset. The details of the datasets are shown in Appendix A.

4.2 Implementation Details

Evaluation Metrics: For MOSI and MOSEI, we report binary classification accuracy (Acc-2), F1 score, seven-class accuracy (Acc-7), mean absolute error (MAE), and correlation (Corr). Acc-2 and F1 are reported in two forms: the first for negative/non-negative (including 0), and the second for negative/positive. Acc 7 shows the percentage of correct predictions made by the model in seven sentiment intervals from -3 to +3. For the SIMS dataset, we report Acc-2, F1, MAE, and Corr. On the CHERMA dataset, F1 score evaluates the model's performance on seven emotion categories. Except for MAE, higher values for all metrics indicate better performance.

Experimental Setup: All models are trained using PyTorch on an NVIDIA RTX A40 with a fixed random seed (1111), and the results are averaged over three experiments. For the MOSI, MOSEI, and SIMS datasets, Adam is used as the optimizer, BERT serves as the backbone, with a learning rate of 5e-6, a batch size of 32, and 100 epochs. The experimental setup for MISR on the CHERMA dataset remains the same as LFMIM (Sun et al., 2023). Appendix B provides detailed of Experimental Setup.

4.3 Baselines

We compare MISR with mainstream and state-of-the-art methods, as follows: TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), TMBL (Huang et al., 2024), CMHFM (Wang et al., 2024a), HyCon (Mai et al., 2022), MMIM (Han et al., 2021), ConKI (Yu et al., 2023), Self-MM (Yu et al., 2021), MulT (Tsai et al., 2019), EFT (Sun et al., 2022), LFT (Sun et al., 2022), MulT (Tsai et al., 2019), PMR (Lv et al., 2021b), LFMIM (Sun et al., 2023), and ALMT (Zhang et al., 2023).

4.4 Comparative Analysis of Experimental.

We analyze the experimental results of MISR and baseline on four datasets in detail. As shown in Tables 1, 2, and 3, MISR outperforms all comparison models on every dataset.

As shown in Table 1, MISR outperforms the comparison models on all metrics for MOSI and MOSEI. Specifically, compared to the SOTA model ALMT (Zhang et al., 2023), MISR achieves an average performance improvement of 1.69% and 1.12%. The performance improvement of MISR is mainly due to its effective use of data distribution discrepancies and sentiment conflicts to optimize the model for learning stable and invariant

Model	Acc-2	F1	MAE	Corr
TFN*	75.27	75.56	0.488	0.496
LMF*	75.36	75.78	0.487	0.502
MulT*	75.62	75.84	0.485	0.504
MISA*	75.49	75.85	0.472	0.542
MAG-BERT*	71.43	63.68	0.553	0.242
Self-MM*	77.37	77.54	0.458	0.535
MMIM*	69.37	58.00	0.607	-
ConKI*	77.94	78.17	0.454	0.542
MISR	**81.53**	**81.91**	**0.425**	**0.597**

Table 2: Performance of models on SIMS. "*" indicates the results are from ConKI (Yu et al., 2023).

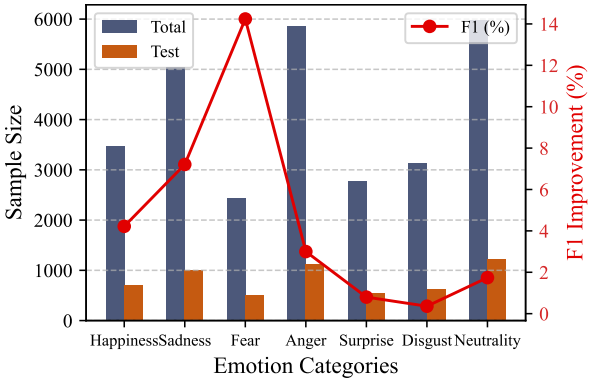

Figure 3: Sample distribution and F1 performance improvement across emotion categories on CHERMA (Compared to LFMIM).

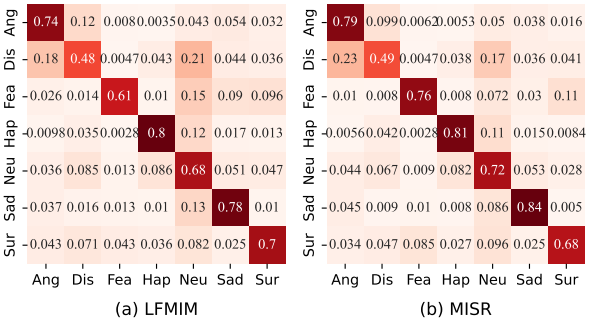

Figure 4: Confusion matrix of emotion categories on CHERMA. Ang, Dis, Fea, Hap, Neu, Sad, and Sur represent Anger, Disgust, Fear, Happiness, Neutrality, Sadness, and Surprise.

sentiment representation. In contrast, existing baselines mainly focus on multimodal fusion or contextual representations, ignoring the impact of dataset distribution imbalance and cross-modal sentiment conflicts on model performance. However, MISR incorporates these unfavorable factors as optimization objectives during training through distribution constraint and invariance constraint, resulting in more robust and effective MSA performance.

As shown in Table 2, we compare the performance of MISR with the baseline model on the SIMS dataset. The results show that MISR outperforms the suboptimal model, ConKI (Yu et al., 2023), with an overall performance improvement of 6.48% across four metrics.

Table 3 presents the performance of different models on the multimodal emotion recognition (MEA) task, including overall accuracy (i.e., overall F1 score) and F1 scores for each emotion category. The results show that MISR outperforms competing models in both overall accuracy and F1 scores for individual categories. Compared to the current SOTA model LFMIM (Sun et al., 2023), MISR achieves a 4.51% improvement in the average F1 score across seven emotion categories.

Fig. 3 shows the distribution of emotion categories in the CHERMA dataset and the F1 performance improvement of MISR over LFMIM. MISR achieves significant performance improvements in recognizing various emotion categories, with notable increases of 7.21% and 14.24% in "Sadness" and "Fear". It is worth noting that the "Fear" category has the fewest samples among the seven emotion categories, yet MISR achieves the highest performance improvement in this category. This further demonstrates that MISR better maintains performance stability across categories when dealing with imbalanced data.

4.5 Visualizing the Impact of Data Imbalance

As shown in Fig. 3, the distribution of the seven emotion categories in the CHERMA dataset is significantly imbalanced. Fig. 4 further presents the confusion matrices for LFMIM and MISR in the seven-category emotion recognition task, facilitating a better analysis of model performance.

As shown in Fig. 4, the predictive recall for the low-frequency emotion category "Fear" is approximately 61% for LFMIM and 76% for MISR. Further analysis reveals that around 15% of the "Fear" samples are misclassified as the high-frequency category "Neutrality" by LFMIM, while the misclassifications of MISR are mainly concentrated in the low-frequency category "Surprise". This indicates a model bias in LFMIM, which tends to misclassify low-frequency samples as high-frequency categories in imbalanced data. In contrast, MISR does not show significant prediction bias in low-frequency categories. However, the recall rate for

Model	Happiness	Sadness	Fear	Anger	Surprise	Disgust	Neutrality	Overall
TFN*	74.91	75.56	66.15	74.41	66.29	43.34	65.60	68.37
LMF*	74.52	75.83	66.73	74.55	65.08	45.70	65.64	68.23
EFT*	74.98	76.88	67.32	74.85	66.73	47.48	64.60	68.72
LFT*	75.07	76.29	66.80	74.88	66.67	47.74	65.97	69.05
MULT*	76.18	76.88	67.36	74.85	68.18	46.96	65.26	69.24
PMR*	75.68	76.46	67.97	75.43	67.37	48.93	66.59	69.53
LFMIM*	76.60	77.83	69.44	75.32	69.83	50.20	68.24	70.54
MISR	**79.83**	**83.44**	**79.33**	**77.58**	**70.39**	**50.38**	**69.43**	**73.75**

Table 3: The comparison with baselines on CHERMA. "*" indicates the results are from LFMIM (Sun et al., 2023).

Method	Acc-2	F1	Acc-7	MAE	Corr	Avg.
MISR	**86.01 / 88.11**	**86.10 / 88.15**	**49.85**	0.671	**0.819**	-
w/o $\mathcal{L}_{NA}$	85.28 / 87.65	85.31 / 87.68	48.10	**0.663**	0.817	0.77% $\downarrow$
w/o $\mathcal{L}_{CMA}$	85.71 / 87.76	85.76 / 87.81	48.83	0.674	0.811	0.71% $\downarrow$
w/o $\mathcal{L}_{DR}$	84.84 / 87.50	85.02 / 87.59	48.54	0.685	0.806	1.46% $\downarrow$
w/o $\mathcal{L}_{IC}$	85.13 / 87.35	85.23 / 87.39	47.86	0.682	0.814	1.42% $\downarrow$

Table 4: Ablation experiments on MOSI.

the "Surprise" category is slightly lower for MISR compared to LFMIM. Further analysis showed that although MISR has a lower recall rate in the "Surprise" category than LFMIM, LFMIM misclassifies more samples from other categories as "Surprise", leading to a lower accuracy. As a result, the overall F1 score for the "Surprise" category is still lower than the MISR.

4.6 Ablation Study

The performance of the MISR model is influenced by four key optimization objectives: normalized alignment $\mathcal{L}_{NA}$, cross-modal alignment $\mathcal{L}_{CMA}$, data reconstruction $\mathcal{L}_{DR}$, and invariance constraint $\mathcal{L}_{IC}$. We evaluate the impact of removing these objectives on MISR using the MOSI dataset.

As shown in Table 4, removing different modules of MISR leads to a performance degradation, indicating that each optimization objective is effective for the task. Further analysis reveals that removing $\mathcal{L}_{DR}$ results in an overall performance degradation of 1.46%, with the largest impact on Acc-7, which decreased by 2.63%. The main reason may be that latent variables z^m serve as the core abstract representation of input data h^m, reflecting the effectiveness and rationality of latent space parameterization. The data reconstruction $\mathcal{L}_{DR}$ constraint ensures that the original data can be recovered as much as possible from the latent space, optimizing the model while preserving effective features. Removin $\mathcal{L}_{IC}$ results in a 1.43% overall performance degradation in MISR, indicating that although the sentiment distribution in the MOSI is relatively balanced, cross-modal sentiment conflicts still exist. Incorporating sentiment conflicts as an optimization objective through invariance learning helps mitigate their adverse impact on the model.

5 Conclusion

This paper proposes the Multimodal Invariant Sentiment Representation Learning (MISR) method, which learns multimodal invariant sentiment representation based on data distribution discrepancies and sentiment conflicts to achieve effective and robust MSA. This paper shows that the imbalance in data distribution provides an optimization direction for distribution constraints, while sentiment conflicts combined with joint representation construct diverse sentiment distribution environments for invariant learning. MISR incorporates unfavorable factors in the data distribution into optimization objectives for model training. Comprehensive experiments across multiple datasets demonstrate that MISR significantly outperforms existing models, providing a new research direction for achieving effective and robust MSA.

Limitations

Although MISR performs well across multiple datasets, it still has some limitations that need to be addressed in future work. First, the tuning of hyperparameters, especially the part related to the loss function, makes it difficult to optimize all evaluation metrics simultaneously. This suggests that more refined optimization strategies are needed to balance different performance demands and further improve the overall performance. Second, although MISR performs well in most emotion categories, there is still room for improvement in classifying the "Disgust" emotion. As a more complex emotional expression, "Disgust" is prone to confusion with other negative emotions (e.g., Anger and Sadness), which affects the classification accuracy and stability of the model. Future research can try to introduce finer-grained emotion feature extraction methods to enhance the model's performance in analyzing complex emotions (e.g., Disgust). Additionally, exploring more diverse datasets and emotional features would help improve the model's robustness and accuracy, further advancing the field of multimodal sentiment analysis.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176084, and Grant 62176083, and in part by the National Key Research and Development Program of China under Grant 2023YFC3604704.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, MA KAILI, Han Yang, Peilin Zhao, Bo Han, and James Cheng. 2023. Pareto invariant risk minimization. In *International Conference on Learning Representations*.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6:169–200.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

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.

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.

Jian Huang, Yanli Ji, Zhen Qin, Yang Yang, and Heng Tao Shen. 2023. Dominant single-modal supplementary fusion (simsuf) for multimodal sentiment analysis. *IEEE Transactions on Multimedia*.

Jiehui Huang, Jun Zhou, Zhenchao Tang, Jiaying Lin, and Calvin Yu-Chian Chen. 2024. Tmbl: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems*, 285:111346.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation

(rex). In *International conference on machine learning*, pages 5815–5826. PMLR.

Zhao-Rong Lai and Weiwen Wang. 2024. Invariant risk minimization is a total variation model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25913–25935. PMLR.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6631–6640.

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.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021a. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2554–2562.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021b. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2554–2562.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14:2276–2289.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. 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 5434–5445, Online. Association for Computational Linguistics.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12966–12978, Toronto, Canada. Association for Computational Linguistics.

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.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2021. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9.

Upendra Singh, Kumar Abhishek, and Hiteshwar Kumar Azad. 2024. A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9):1–38.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 658–670, Toronto, Canada. Association for Computational Linguistics.

Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 15–23.

Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. 2022. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision*, pages 709–726. Springer.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. 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.

Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, et al. 2024a. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management*, 61(3):103675.

Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. 2024b. Dissecting the failure of invariant learning on graphs. In *Advances in Neural Information Processing Systems*, volume 37, pages 80383–80438. Curran Associates, Inc.

Jiayun Wu, Jiashuo Liu, Peng Cui, and Steven Z. Wu. 2024. Bridging multicalibration and out-of-distribution generalization beyond covariate shift. In *Advances in Neural Information Processing Systems*, volume 37, pages 73036–73078. Curran Associates, Inc.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.

Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion recognition for multiple context awareness. In *European conference on computer vision*, pages 144–162. Springer.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. ConKI: Contrastive knowledge injection for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13610–13624, Toronto, Canada. Association for Computational Linguistics.

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.

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*.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore. Association for Computational Linguistics.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. 2024. Kebr: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 5732–5741, New York, NY, USA. Association for Computing Machinery.

A Datasets

We evaluate the performance of the MISR model through two tasks: Multimodal Sentiment Analysis (MSA) and Multimodal Emotion Recognition (MER).

MSA: MSA aims to analyze the sentiments expressed by people in a video. For the MSA task, we evaluate the performance of MISR using three widely recognized benchmark datasets: MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and SIMS (Yu et al., 2020). The MOSI dataset contains 2,199 annotated video clips, in which speaker express opinions on topics such as movies, drawn from 93 YouTube videos. MOSEI, an extended version of MOSI, contains 22,856 annotated video clips covering 250 different topics. Each clip in both datasets is labeled with sentiment intensity, ranging from -3 (strongly negative) to +3 (strongly positive). SIMS is a MSA dataset containing 2,281 video clips. Every sample is annotated with one multimodal label and three unimodal labels, with sentiment scores ranging from -1 to +1.

MER: MER focuses on classifying verbal emotions in videos into multiple emotion categories. We conducted experimental analyses using the emotion recognition dataset CHERMA (Sun et al., 2023). CHERMA contains 28,717 video clips from different media, categorized according to Ekman's six basic emotions (Ekman, 1992) (i.e., happiness, sadness, fear, anger, surprise, and disgust), as well as neutrality. Each video clip was labeled with three unimodal labels and one multimodal label. The dataset was divided into training, validation, and test sets in the ratio of 6:2:2.

Table 5 shows the statistical information of all datasets.

Dataset	Speaker	Video clip	Train	Valid	Test	Language
MOSI	93	2199	1284	229	686	English
MOSEI	1000	22856	16326	1871	4659	English
SIMS	474	2281	1368	456	457	Chinese
CHERMA	-	28717	17230	5743	5744	Chinese

Table 5: Dataset statistics.

Hyper-parameter	MOSI	MOSEI	SIMS	CHERMA
d_t	768&1024	768&1024	768&1024	1024
d_a	5	74	33	1024
d_v	20	35	709	2048
λ_1, λ_2, λ_3	1,0.4,1	1,0.4,1	1,0.5,1	1,0.5,1
Batch size	32	32	32	24
Epoch	100	100	100	50
Optimizer	Adam	Adam	Adam	SGD
Vector Length T	50	50	39	80
Learning rate of BERT	5e-6	5e-6	5e-6	5e-4
Learning rate of others	1e-4	1e-4	1e-4	5e-4
Fully connected layer	128	128	128	1024

Table 6: Hyper-parameters setting.

Backbone	Parameters	Time / Epoch
BERT-base	115 M	8 s
BERT-large	341 M	17 s

Table 7: Computational overhead.

B Hyper-parameters setting

Table 6 provides a detailed of hyper-parameters setting.

C Computational Overhead

Table 7 presents the computational overhead analysis of the MISR model on the MOSI (Zadeh et al., 2016) dataset. As shown, the use of BERT-base, which has fewer parameters, results in faster runtime, but with a potential trade-off in performance. On the other hand, BAER-large, with more parameters, takes longer to run but may achieve better performance due to its increased model capacity. This highlights the balance between computational efficiency and model performance when selecting model configurations.

D Rationale for Multimodal IRM

In single-modal tasks, IRM improves the model's generalization ability across different data distributions by learning invariant features that are stably related to the target labels. "Invariance" does not mean that the data itself or the feature distribution remains completely unchanged; rather, it refers to the model learning features that consistently maintain a stable association with the label across different environments. For example, in image classification, IRM encourages the model to focus on object shapes rather than unstable spurious features like background.

When extending to multimodal sentiment analysis (MSA), the imbalance in single-modal data distribution and cross-modal sentiment conflicts make direct modeling more challenging. However, multimodal fusion can often compensate for the instability of individual modalities, making multimodal sentiment representations invariant to unimodal sentiment polarity and data distribution. For instance, a single modality may be affected by noise or biases (e.g., text sentiment classification may be skewed due to sarcasm), but multimodal fusion enhances robustness and mitigates such effects. Moreover, from the IRM perspective, modality conflicts and sentiment distribution imbalance in MSA naturally form different data distribution environments, providing a reasonable optimization scenario for invariant risk minimization.

Based on this, we propose I Multimodal Invariant Sentiment Representation Learning (MISR), which draws inspiration from IRM to optimize in-

No.	IR	IC	Text	Audio	Visual	Acc-2	F1	Acc-7	MAE	Corr
N1	H^u	✓	×	×	×	86.01 / 88.11	86.10 / 88.15	49.85	0.671	0.819
N2	H^u	✓	✓	×	×	85.88 / 87.95	85.91 / 87.91	49.55	0.678	0.804
N3	H^u	×	✓	×	×	85.77 / 87.66	85.74 / 87.62	48.73	0.684	0.799
N4	H^u	✓	×	✓	×	85.83 / 87.88	85.92 / 87.76	48.70	0.681	0.806
N5	H^u	×	×	✓	×	85.40 / 87.33	85.56 / 87.40	48.21	0.690	0.801

Table 8: Ablation analysis under different environments. Partial examples, where IR and IC represent invariant representations and invariant constraints, respectively.

variant sentiment representations in a multimodal setting. MISR first learns stable cross-modal joint sentiment representations, then, under IRM constraints, further refines invariant features from different modality combinations and distribution environments, thereby improving the model's generalization ability in MSA tasks.

E Experimental Validation of Extending IRM to Multimodal

In the ablation study, we provide multiple distribution environments built in Section 3.4 and compare the model performance with and without the invariance constraint. As shown in Table 8.

A comparison between N2 and N3 shows better performance under the learned joint invariant representation IR with invariant constraints in the text environment. Similarly, N2 and N4 outperform N3 and N5, where invariant constraints are removed. This suggests that invariant constraints are effective, supporting the hypothesis. Additional environment tests are in the supplement.