InfoEnh: Towards Multimodal Sentiment Analysis via Information Bottleneck Filter and Optimal Transport Alignment

Yifeng Xie[1,†] **Zhihong Zhu**[2,†] **Xuan Lu**[1] **Zhiqi Huang**[2,*] **Haoran Xiong**[1]
[1]Guangdong University of Technology, [2]Peking University
evfxie@gmail.com, zhihongzhu@stu.pku.edu.cn,
{3221006905,3121006856}@mail2.gdut.edu.cn, zhiqihuang@pku.edu.cn

Abstract

In recent years, Multimodal Sentiment Analysis (MSA) leveraging deep learning has demonstrated exceptional performance in a wide range of domains. Its success lies in effectively utilizing information from multiple modalities to analyze sentiments. Despite these advancements, MSA is confronted with two significant challenges. Firstly, each modality often has a surplus of unimportance data, which can overshadow the essential information. Secondly, the crucial cues for sentiment analysis may conflict across different modalities, thereby complicating the analysis process. These issues have a certain impact on the model's effectiveness in MSA tasks. To address these challenges, this paper introduces a novel method tailored for MSA, termed InfoEnh. This approach utilizes a masking technique as the bottleneck for information filtering, simultaneously maximizing mutual information to retain crucial data. Furthermore, the method integrates all modalities into a common feature space via domain adaptation, which is enhanced by the application of optimal transport. Extensive experiments conducted on two benchmark MSA datasets demonstrate the effectiveness of our proposed approach. Further analyzes indicate significant improvements over the baselines.

Keywords: multimodal learning, information bottleneck, optimal transport

1. Introduction

Multi-modal learning has gained widespread popularity, emerging as a breakthrough in various research domains and industries. At the same time, there has been a growing research interest in the field of multi-modal sentiment analysis (MSA), a prominent task within multi-modal learning. MSA is a crucial component for understanding human expression and interaction, encompassing various aspects like user engagement (Spiess and Schuldt, 2022), personalized recommendations (Chen et al., 2019b), conversational systems (Huang et al., 2023; Xie et al., 2023; Zhu et al., 2023), content moderation (Yuan et al., 2024), risk assessment (Ang and Lim, 2022) and more. Its ability to accurately identify the sentiments of individuals involved can greatly improve the performance of artificial intelligence products.

Generally speaking, different modalities present distinct information, resulting in a wealth of data that is significantly more robust than what any single modality can provide. For example, consider a word in a text with multiple meanings, each conveying different emotions depending on the context (Zadeh et al., 2017). In such instance, the role of additional modalities becomes particularly vital, as they provide essential contextual cues that help discern the ultimate emotion. In more intricate scenarios, like sarcasm (Castro et al., 2019), a single modality proves insufficient to accurately encapsulate the underlying sentiment. Therefore, the key to overcoming the current challenges in MSA lies in the successful integration of significant information from various modalities. Numerous works (Wang et al., 2022; Yadav and Vishwakarma, 2023) have been proposed, resulting in significant research achievements and successes.

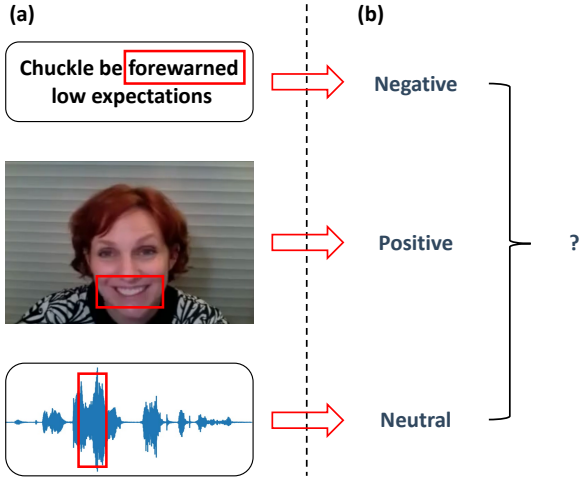

Figure 1: An example of Multimodal Sentiment Analysis. The red rectangular boxes represent the most critical features for sentiment in this modality. The sample is selected from MOSI dataset.

While significant progress has been made, we have identified two primary challenges with existing methods: (1) **Redundant Information.** Recent studies have highlighted the presence of significant amounts of irrelevant or redundant informa-

[†]Equal Contribution.
[*]Corresponding author.

LREC-COLING 2024, pages 9073–9083

tion in the modalities (Wu et al., 2023; Mao et al., 2023). As shown in Figure 1(a), it is possible to derive emotions expressed by the modalities from the most critical features, rather than processing all features within the model. Processing numerous irrelevant or useless features can negatively impact the model's efficiency. Regrettably, existing models often overlook this aspect. (2) **Contradictory Modality Sentiment.** As shown in Figure 1(b), different modalities may express varying sentiments after processing. In other words, features from different modalities may be ambiguous and conflicting, which divergence can lead to confusion in the model's final judgment.

Motivated by the above observation, we propose a novelty module termed InfoEnh to improve the performance of MSA. InfoEnh initiates by identifying the top-k features based on their weights, followed by an element-wise multiplication with a generated soft mask. These steps act as an information bottleneck, effectively filtering numerous non-essential or irrelevant features. To ensure that key features are not filtered out, we maximize the mutual information between the features before and after this filtering process as a regulation measure. Moreover, humans excel at distinguishing between different modalities of information, whereas models often struggle with this task (Xu et al., 2021a; Wei et al., 2023; Zeng et al., 2023). This significant capability gap prompted our exploration of a potential solution involving optimal transport (OT). OT offers a method to quantify dissimilarity between two distributions. Our hypothesis suggests that leveraging OT can effectively integrate diverse unimodal features into a shared subspace, bridging the distribution gap and thus creating a consistent and comprehensive representation in the unified multimodal space.

In summary, the major contributions of this paper can be summarized as:

- We present a novel method for MSA designed for easy and seamless integration with existing MSA frameworks. This innovative method utilizing information bottleneck prioritizes the removal of irrelevant features while safeguarding essential and significant data.
- Unlike previous studies in MSA that primarily focused on multimodal fusion, our research is centered on multimodal representation. Our proposed InfoEnh leverages the power of optimal transport. This integration equips the model to handle even more intricate scenarios, allowing the model to train the heterogeneous data of multiple modalities.
- Experimental results demonstrate the InfoEnh achieves excellent performance on multimodal benchmarks based on the improvement of each baseline. Furthermore, we offer an in-depth analysis supported by lots of results to illustrate the superiority of our proposed method.

2. Related Work

2.1. Multimodal Sentiment Analysis

The MSA task is inherently challenging, requiring the seamless integration of information from diverse modalities, including text, visual, audio and more. It involves accurately capturing the intricate interplay between these modalities, and acknowledging the multimodal nature of human communication. Integrating diverse data types can lead to a nuanced and precise sentiment analysis. This complexity is further compounded by the potential for conflicting yet complementary information, rendering the task particularly challenging.

In essence, MSA is a subset of multi-modal learning. The fusion on MSA can be broadly categorized into two approaches (Snoek et al., 2005; Gadzicki et al., 2020): (1) Early fusion. In this approach, different modalities' inputs are combined at a shallow layer (Poria et al., 2016; Wang et al., 2017; Guo et al., 2022), merging features from individual modalities into a shared model input parameter space. (2) Late fusion. Data from different modes undergo modeling and feature extraction using distinct network structures. The features extracted from these diverse modalities are combined before the final output (Nojavanasghari et al., 2016; Qian et al., 2023; Yu et al., 2023), unifying them into a shared feature space. Various fusion approaches in MSA are well-developed and have shown significant progress.

However, advancing MSA requires more than just focusing on the fusion module. It also necessitates the extraction of the most appropriate features. Our work utilizes information bottleneck to filter features and then harmonizes the feature spaces of different modalities using optimal transport. This strategy enhances the efficacy of MSA.

2.2. Information Bottleneck

The Information Bottleneck (IB) method (Tishby et al., 2000) provides a strategy for obtaining an intermediate variable Z that extracts the core information from the input embedding X, aligning it more closely with the true output Y. The method formulates an optimization challenge aimed at compressing X while retaining maximal information about Y. The objective is to maximize the mutual information between the compressed representation Z and Y, while imposing a constraint to limit the mutual information between Z and X.

Actually, the IB method has been widely applied to various problems across different domains. More recently, it has been utilized in the field of deep learning, offering insights into the training dynamics and generalization capabilities of deep neural networks. (Tishby and Zaslavsky, 2015) explored the IB method within the context of deep learning, suggesting that deep networks perform an IB-like compression process due to their layered architecture. This concept was further developed by (Shwartz-Ziv and Tishby, 2017), which empirically demonstrated the phases of compression and fitting during the training of deep models. Despite its extensive applicability, the Information Bottleneck method does have its limitations. One significant challenge is the computational complexity involved in evaluating mutual information for continuous and high-dimensional variables. This issue has been tackled using variational approximations (Alemi et al., 2017). As demonstrated, the IB method has yielded numerous groundbreaking insights and contributions.

In this work, we leverage the foundational principles of the IB method to enhance the performance of MSA task. This is achieved by filtering out unimportant features and focusing on the most significant ones, while also maximizing the mutual information between the compressed and input variables. Our goal is to efficiently compress the information while preserving the most vital feature.

2.3. Optimal Transport

Optimal Transport (OT) was initially used to solve transportation problems (Monge, 1781), where a certain amount of resources needs to be transported from one location to another, to find a measurable map to minimize the cost of transportation. To put it mathematically, OT focuses on optimizing over transportation plan, which is a probability measure to preserve their respective quantity (Kantorovich, 2006; Villani, 2021). In the field of machine learning, OT has gained prominence for its ability to define a geometrically meaningful distance metric between probability distributions (Zhu et al., 2024), known as the Wasserstein distance (Vaserstein, 1969).

OT has since been utilized in various NLP tasks (Chen et al., 2019a; Cao and Zhang, 2022; Bhardwaj et al., 2022; Cheng et al., 2024) and multi-modal learning (Cao et al., 2022; Pramanick et al., 2022; Chen et al., 2023) due to its ability to establish a reasonable correspondence between two distribution sets. It aligns the embeddings, creates a common metric space between different embeddings, and thereby facilitates transfer learning (Alqahtani et al., 2021). Furthermore, OT measures the distance between probability distributions, enhancing machine translation models for better source-target language similarity (Xu et al., 2021b). It can also generate realistic text by mapping training data probability distributions to desired output distributions (Nouri, 2022). In addition, it is widely applied in numerous theoretical and practical contexts.

In this work, we utilize OT for domain adaptation to multimodal embeddings. We leverage the theory of optimal transport to align embeddings and address the challenges in MSA.

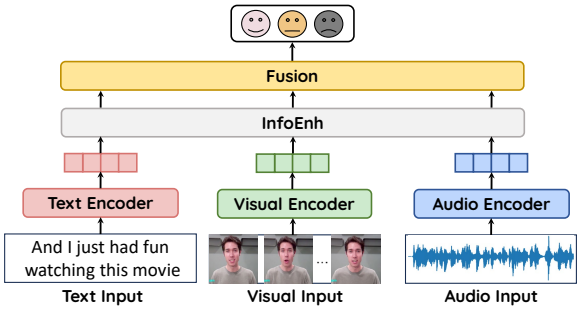

Figure 2: The overall training process for MSA.

3. Preliminaries

3.1. Problem Definition

Given the inputs **x**, each sample comprises three modalities: **t**ext (t), **v**isual (v) and **a**udio (a). When these multimodal inputs are processed by the model $f(\cdot)$, it produces an output $\mathbf{y} \in [-3, 3]$ that predicts the sentiment associated with each input. The training process is defined by $\mathbf{y} = f(t, v, a|\Theta)$, where Θ represents the model parameters. Further details on the model's training procedure are elaborated in Section 3.2. In essence, this task is a regression problem, where the output score indicates sentiment: zero represents neutral sentiment, positive scores indicate positive sentiment and negative scores suggest negative sentiment.

3.2. Model Training

Upon entering the encoder, inputs from each modality are transformed into their respective embeddings. These embeddings are then fused to produce a predicted value. Following the previous works (Hazarika et al., 2020; Zhang et al., 2023), the model employs both predicted and actual values to calculate either the mean squared error loss or the cross-entropy loss, to minimize the loss to optimize the model. Most existing models prioritize feature integration, emphasizing the design of sophisticated fusion modules.

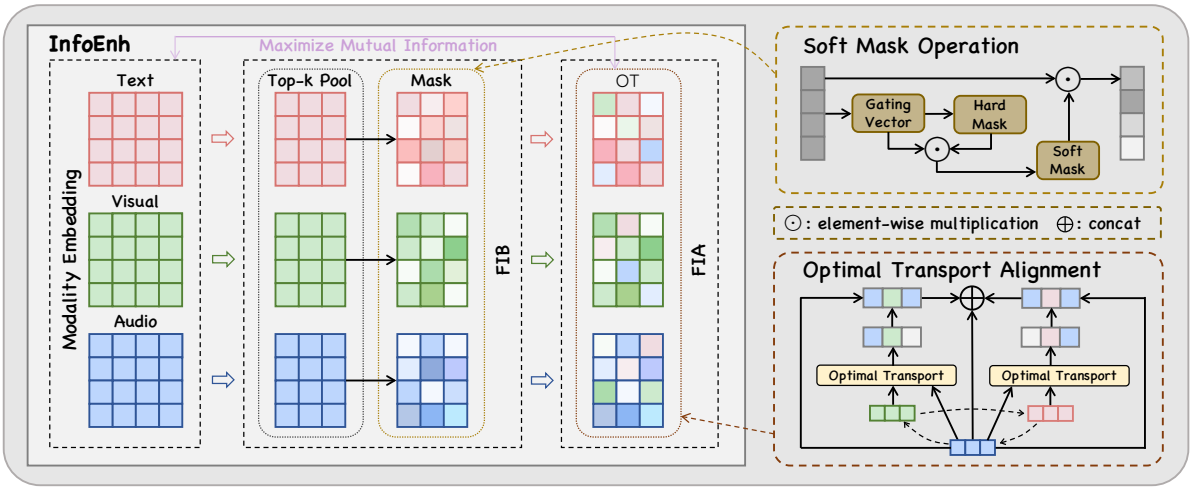

Figure 3: Illustration of InfoEnh: The embeddings are processed through FIB (§4.1) and FIA (§4.2).

4. Methodology

As mentioned above and illustrated in Figure 2, our model for MSA resembles most existing models with one key distinction: after extracting modality features, we input the embeddings into our proposed InfoEnh component, followed by the fusion module. Existing fusions are already well-developed. Therefore, our proposed InfoEnh focuses on obtaining the most relevant and consistent features for incorporation into the architecture.

As shown in Figure 3, InfoEnh is composed of two components: **F**eature **I**nformation **B**ottleneck (FIB) and **F**eature **I**nformation **A**lignment (FIA). FIB is designed to identify and select the most crucial features, while FIA seeks to harmonize the features from each modality within a unified distribution space. Further details on these components can be found in Section 4.1 and 4.2.

4.1. Feature Information Bottleneck

Upon obtaining the embeddings from each modality, we often encounter an abundance of irrelevant and non-essential features. However, for sentiment analysis, only the most pivotal features are required in most cases. To address this, we utilize the Feature Information Bottleneck (FIB). In the following, we provide a detailed description.

Initially, the input embedding $\mathbf{E} \in \mathbb{R}^{l_i \times d_i}$ undergoes processing by a feed-forward neural network (FNN) layer, followed by a pooling layer. In the pooling layer, the top-k feature weights are selected based on their highest values, thereby identifying these features as particularly significant. It is important to highlight that after extracting the features from different modalities, the dimensions may not be uniform across models. This step ensures the standardization of dimensions to k.

Subsequently, the new representation $\mathbf{e} \in \mathbb{R}^{k \times d}$ undergoes an element-wise multiplication with a soft mask $\mathbf{m} \in \mathbb{R}^{k \times d}$ to selectively filter out non-essential or irrelevant feature information. The soft mask is obtained from the element-wise product of a gating vector $\mathbf{g}$ and a hard mask $\mathbf{m}_h$. The gating vector is produced from the embedding matrix after it has been processed through a FNN layer and a nonlinear activation function σ, such as sigmoid function (Han and Moraga, 1995), which assigns a value between 0 and 1 to each feature.And the assignment of values in the hard mask depends on the gating vector: if a value within the gating vector surpasses the threshold (set here at 0.5), the corresponding hard mask value is assigned as 1, thereby preserving the feature. Conversely, if the threshold is not met, the hard mask value is set to 0, effectively eliminating the feature. Thus, we obtain the filtered embedding, denoted as $\mathbf{z} \in \mathbb{R}^{k \times d}$. The formulas as follows:

$$\mathbf{g} = \sigma(\mathbf{w} \cdot \mathbf{e} + \mathbf{b}) \quad (1)$$

$$\mathbf{m} = \mathbf{m}_h \odot \mathbf{g} \quad (2)$$

$$\mathbf{z} = \mathbf{e} \odot \mathbf{m} \quad (3)$$

where $\mathbf{w}$ and $\mathbf{b}$ are the parameters of the FFN layer, $\odot$ denotes the element-wise multiplication.

The information bottleneck principle is designed to optimize the representation, aiming to maximize the mutual information pertaining to the target output. Concurrently, it seeks to minimize the retention of input information. This approach ensures that the representation captures the most relevant aspects of the input embedding. However, the new feature $\mathbf{z}$ may result in the loss of key information. To address this, we aim to effectively eliminate redundant features while preserving significant mutual information between the new and original features, thus retaining essential information. To accomplish this objective, we use a similarity function

$sim(\cdot)$ to produce the InfoNEC loss (van den Oord et al., 2018; Wu et al., 2023).

$$sim(\mathbf{e}, \mathbf{z}) = \exp(\frac{\mathbf{e}}{||\mathbf{e}||_2} \odot \frac{\mathbf{z}}{||\mathbf{z}||_2}) \quad (4)$$

$$\mathcal{L}_i = \mathbb{E}_{(e,z)} \left[-\sum_{i=1}^{n} log \frac{\exp(sim(\mathbf{e}^i, \mathbf{z}^i))}{\sum_{k=1}^{m} \exp(sim(\mathbf{e}_k^-, \mathbf{z}^i))} \right], \ i \in \{t, v, a\} \quad (5)$$

$$\mathcal{L}_{FIB} = \alpha(\mathcal{L}_t + \mathcal{L}_v + \mathcal{L}_a) \quad (6)$$

where $||\cdot||_2$ denotes the Euclidean norm, n represents the number of samples in a batch, $\mathbf{e}^-$ denotes the negative samples, indicating that their sentiment differs from the sentiment of the sample $\mathbf{z}^i$. The expected value $\mathbb{E}$ refers to the mean value under this probability distribution, and α is a hypermeter. In summary, we introduce an additional loss function $\mathcal{L}_{FIB}$, generated by each modality $\mathcal{L}_i$, which serves to augment the task loss throughout the training process.

4.2. Feature Information Alignment

Although key features are extracted for different modalities, the embeddings often reside in disparate heterogeneous spaces. The alignment process is crucial for mapping features from varied distributions into a unified feature space. This enables the model to more effectively discern the relationships between modalities, distinguishing between features that are consistent and those that are contradictory, to inform the final sentiment analysis.

Optimal transport (OT) offers a method for measuring and aligning the feature distributions across these diverse modalities. As shown in Figure 4, OT moves different modality embeddings to a unified aligned space using an optimal transport plan. This strategy effectively addresses spatial heterogeneity by minimizing the Wasserstein distance between the distributions.

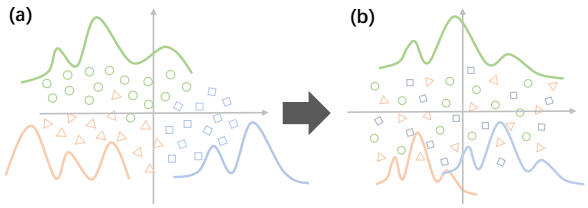

Figure 4: An illustration of the optimal transport process: Different colors represent different modalities. Lines represent distributions, while markers represent the embeddings.

Specifically, we formulate the fusion procedure of multimodal knowledge as an optimal transport problem. Here, we only focus on the discrete situation, which is more related to our framework. Suppose the distribution $\mathbf{U}$ corresponds to one embedding, and the distribution $\mathbf{V}$ corresponds to another embedding. It's important to note that the number of things transported in a distribution cannot exceed the sum of the quantities of the original distribution. In mathematical terms, we represent these distributions as $\mathbf{U} = \sum_i u_i \delta_i$ and $\mathbf{V} = \sum_i v_i \delta_i$, where δ_i is the Dirac function. In our work, we simplify the process by setting the weights $u_i = 1/m$ and $v_i = 1/n$, where m and n represent the length of the embeddings. Next, we aim to compute the optimal transport plan $\mathbf{P}_{ot}$, which seeks to obtain an optimal transport matrix to minimize the transport cost $\mathbf{C}$. In our work, we set $\mathbf{C} = 1 - \mathbf{U}^T\mathbf{V}$.

Algorithm 1 Fusion of two modality embeddings using optimal transport.

Initialize: Entropy parameter λ, the maximum number N of iterations.
Input: One modality embedding $\mathbf{E}_i \in \mathbb{R}^{k \times d}$, other modality embedding $\mathbf{E}_j \in \mathbb{R}^{k \times d}$.
Output: Unified embedding $\mathbf{E}_{i \to j} \in \mathbb{R}^{k \times d}$.

Calculate cost metric $\mathbf{C} = 1 - \mathbf{E}_i^T \mathbf{E}_j$.
Initialize $\mathbf{E}_j^{(0)} = 1, \delta = 0.01, \Delta = \infty$.
In every loop, calculate the OT matrix.
for $k_t = 1, 2, ..., N$ **do**
  Update $\mathbf{E}_i^{(k_t)} = \mathbf{E}_i / (\exp(-\lambda \mathbf{C}) \mathbf{E}_j^{(k_t - 1)})$
  Update $\mathbf{E}_j^{(k_t)} = \mathbf{E}_j / (\exp(-\lambda \mathbf{C}) \mathbf{E}_i^{(k_t - 1)})$
  Update $\Delta = \frac{1}{n} \sum |\mathbf{E}_j^{(k_t)} - \mathbf{E}_j^{(k_t - 1)}|$
  if $\Delta < \delta$ **then**
    break.
  end if
end for
Obtain the OT matrix:
$\mathbf{P}^* = diag(\mathbf{U}^k) \exp(-\lambda \mathbf{C}) diag(\mathbf{V}^k)$.
Calculate the OT plan $\mathbf{P}_{OT} = < \mathbf{P}^*, \mathbf{C} >$.
Obtain **unified embedding** $\mathbf{E}_{i \to j} = \mathbf{E}_i \mathbf{P}_{OT}^T$

$$\mathbf{Q}(\mathbf{U}, \mathbf{V}) = \{\mathbb{R}_+ | \mathbf{P} 1_m = \mathbf{U}, \mathbf{P}^T 1_n = \mathbf{V}\} \quad (7)$$

$$\mathbf{P}_{ot}(\mathbf{U}, \mathbf{V} | \mathbf{C}) = \inf_{P \in Q(\mathbf{U}, \mathbf{V})} < \mathbf{P}, \mathbf{C} >_F \quad (8)$$

$$< \mathbf{P}, \mathbf{C} >_F = \sum_{i,j} \mathbf{P}_{ij} \mathbf{C}_{ij} \quad (9)$$

The formulas for the optimal transport plan $\mathbf{P}_{ot}$ are defined using the Frobenius inner product denoted as $< , >_F$, and the feasible set $\mathbf{U}$ satisfies all the conditions for transport. In other words, the OT problem aims to find the optimal transport plan subject to the constraints of the feasible set.

The above represents a constrained linear programming problem, and direct optimization of the objectives is usually time-consuming. Therefore, we will introduce the Sinkhorn algorithm (Cuturi, 2013) and incorporate it into the process. Firstly,

Model	MOSI					MOSEI				
	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)
FDMER†	0.724	0.788	44.10	- /84.60	- / 84.70	0.536	0.773	54.10	- / 86.19	- / 85.80
DBF†	0.693	0.801	44.80	85.10 / 86.90	85.10 / 86.90	0.523	0.772	54.20	84.30 / 86.40	84.80 / 86.20
ALMT†	0.683	0.805	49.42	84.55 / 86.43	84.57 / 86.47	0.526	0.779	54.28	84.78 / 86.79	85.19 / 86.86
MISA‡	0.796	0.766	42.51	80.49 / 81.88	80.47 / 81.98	0.571	0.723	52.15	82.54 / 84.18	82.54 / 83.86
MISA + InfoEnh	**0.782**	**0.772**	**43.26**	**80.97 / 82.12**	**81.02 / 82.18**	**0.559**	**0.735**	**52.75**	**82.98 / 84.63**	**83.02 / 84.68**
self-MM‡	0.720	0.789	45.68	82.33 / 84.75	82.71 / 84.86	0.536	0.758	53.45	82.49 / 84.88	82.51 / 84.91
self-MM + InfoEnh	**0.709**	**0.802**	**46.12**	**84.40 / 85.65**	**84.43 / 85.72**	**0.530**	**0.764**	**53.82**	**82.98 / 85.20**	**83.01 / 85.24**
MMIM‡	0.708	0.796	46.25	82.81 / 84.95	82.97 / 85.05	0.532	0.765	53.93	82.29 / 85.78	82.38 / 85.86
MMIM + InfoEnh	**0.698**	**0.808**	**46.77**	**84.37 / 85.49**	**84.42 / 85.58**	**0.524**	**0.776**	**54.16**	**83.27 / 86.24**	**83.36 / 86.40**
ConFEDE‡	0.695	0.806	48.62	84.43 / 86.26	84.52 / 86.32	0.528	0.778	54.20	84.48 / 86.56	84.60 / 86.72
ConFEDE + InfoEnh	**0.683**	0.805	**49.25**	**84.57 / 86.65**	**84.60 / 86.74**	**0.520**	**0.785**	**55.38**	**84.78 / 86.98**	**84.82 / 87.01**

Table 1: Main results on two benchmark datasets. The "Acc-2" value corresponds to "negative/non-negative", and the "F1" value corresponds to "negative/positive". Results marked with "†" indicate that the code has not been released. Results marked with "‡" represent we re-implemented models, which achieved statistical improvements over the baselines with $p < 0.05$. Results highlighted in **bold** signify improvements over the baselines.

$\mathbf{P}_{ot}$ is introduced entropy regularization h with a hyperparameter λ, which makes the feasible area of the original problem smoother:

$$\mathbf{P}_{ot,\lambda}(\mathbf{U},\mathbf{V}|\mathbf{C}) = \min_{\mathbf{P}} < \mathbf{P}, \mathbf{C} >_F - \frac{1}{\lambda} h(\mathbf{P}) \quad (10)$$

By taking the partial derivative of the Lagrange function, we obtain an approximate optimization solution with a reduced number of loop iterations, significantly reducing the computational cost. In our proposed framework, when we combine the definitions mentioned earlier, we will obtain an optimal transportation plan $\mathbf{P}^*$:

$$\mathbf{P}^* = diag(\mathbf{U}^{(k)}) \exp(-\lambda \mathbf{C}) diag(\mathbf{V}^{(k)}) \quad (11)$$

where k denotes the iteration and in each iteration to solve: $\mathbf{U}^{(k)} = \mathbf{U}/(\exp(-\lambda\mathbf{C})\mathbf{V}^{(k-1)})$ and $\mathbf{V}^{(k)} = \mathbf{V}/(\exp(-\lambda\mathbf{C})^T\mathbf{U}^{(k)})$.

To align the embeddings of the three modalities in the same space while retaining modality-specific information, each modality should adapt to the domains of the other two modalities.

We firstly obtain the transported feature embeddings using OT method, as demonstrated in Algorithm 1. These embeddings are denoted as $\mathbf{E}_{i\to j} \in \mathbb{R}^{k\times d}$ and represent the features transported from modality i to modality j. Subsequently, we obtain a new feature embedding for each modality $\mathbf{E}_i^{'} \in \mathbb{R}^{k\times 3d}$ by concatenating the original modality embedding with the two transported embeddings after domain adaptation. This concludes the optimization phase of the representations, which are then utilized for fusion in the previous model.

$$\mathbf{E}_{i\to j} = \mathbf{E}_i \mathbf{P}_{OT}^T(i,j), \;\; i \neq j, \;\; i,j \in \{t,v,a\} \quad (12)$$

$$\mathbf{E}_i^{'} = \text{concat}(\mathbf{E}_i, \mathbf{E}_{i\to j}, \mathbf{E}_{i\to k}), \; i \neq j, j \neq k, i \neq k \quad (13)$$

5. Experiment

5.1. Experiment Setup

Datasets. The experiments were conducted on two publicly available multimodal sentiment analysis datasets: MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018). MOSI consists of a collection of YouTube monologues where speakers share their opinions on various topics, including movies. It contains 93 videos from 89 unique speakers, totaling 2,199 movie-related samples. These samples are categorized into 1,284 training samples, 229 validation samples, and 686 test samples. MOSEI extends MOSI by providing a more extensive collection with 23,453 video clips from 1,000 unique speakers, thus offering a more comprehensive resource for analysis.

Evaluation Metrics. Following the previous works (Hazarika et al., 2020; Han et al., 2021; Yu et al., 2023), we utilize the following metrics to evaluate the performance of our model: Mean Absolute Error (MAE), Correlation of the model's prediction with human (Corr), Binary Classification Accuracy (Acc-2), Seven Classification Accuracy (Acc-7), and F1-score (F1). By employing these metrics, we aim to provide a thorough and nuanced assessment of the performance.

Baselines. To comprehensively validate the performance of our InfoEnh, we make a comparison with several advanced and state-of-the-art methods, including: FDMER (Yang et al., 2022), DBF (Wu et al., 2023), ALMT (Zhang et al., 2023), MISA (Hazarika et al., 2020), self-MM (Yu et al., 2021), MMIM (Han et al., 2021) and ConFEDE (Yang et al., 2023).

Model	MOSI					MOSEI				
	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)
self-MM + InfoEnh	**0.709**	**0.802**	**46.12**	**84.40 / 85.65**	**84.43 / 85.72**	**0.530**	**0.764**	**53.82**	**82.98 / 85.20**	**83.01 / 85.24**
w/o FIA	0.712	0.794	46.01	83.98 / 85.10	84.02 / 85.14	0.533	0.763	53.66	82.72 / 84.92	82.78 / 84.95
w/o FIB	0.716	0.792	45.83	83.42 / 84.96	83.46 / 84.99	0.533	0.760	53.58	82.60 / 84.90	82.64 / 84.96
w/o InfoEnh	0.720	0.789	45.68	82.33 / 84.75	82.71 / 84.86	0.536	0.758	53.45	82.49 / 84.88	82.51 / 84.91
ConFEDE + InfoEnh	**0.683**	0.805	**49.25**	**84.57 / 86.65**	**84.60 / 86.74**	**0.520**	**0.785**	**55.38**	**84.78 / 86.98**	**84.82 / 87.01**
w/o FIA	0.688	0.804	49.04	84.52 / 86.52	84.58 / 86.55	0.522	0.785	54.68	84.71 / 86.85	84.75 / 86.92
w/o FIB	0.692	0.800	48.78	84.46 / 86.32	84.50 / 86.36	0.528	0.780	54.32	84.58 / 86.72	84.66 / 86.79
w/o InfoEnh	0.695	0.806	48.62	84.43 / 86.26	84.52 / 86.32	0.528	0.778	54.20	84.48 / 86.56	84.60 / 86.72

Table 2: Experiment results of ablation study for each component across different datasets.

Methods	*w/o* InfoEnh		*w/* InfoEnh	
	Acc-2 (↑)	F1 (↑)	Acc-2 (↑)	F1 (↑)
ConFEDE	84.48	84.60	84.78	84.82
w/o Audio	83.21	83.83	84.04	84.16
w/o Visual	81.83	82.33	82.65	83.68
w/o Text	55.23	54.45	66.75	64.80

Table 3: Comparison of accuracy for missing modalities on the MOSEI dataset.

Implementation Details. The reproducible baseline models and the novel InfoEnh we propose are implemented using the PyTorch framework (Paszke et al., 2017). All experiments were conducted on a single Intel Xeon(R) CPU equipped with an NVIDIA RTX 3080 Ti GPU. The hyperparameter α, which adjusts the loss function of FIB, is set to 0.5. In the interest of a fair and equitable comparison, we adhered to the training configurations detailed in the original publications for each model, including the choice of loss functions, batch sizes, and learning rate schedules. This approach ensures that the performance of our proposed model is benchmarked against established methods under identical conditions.

5.2. Main Results

Table 1 presents the baseline results alongside those achieved by incorporating InfoEnh, highlighting improvements on the MOSI and MOSEI datasets. The detailed examination of the outcomes yields several observations:

A pivotal observation emerges from the Mean Absolute Error (MAE) comparison, where InfoEnh demonstrates superior performance over the existing baselines. This indicates that InfoEnh is particularly effective in deriving meaningful representations for MSA. Moreover, InfoEnh achieves impressive improvements about Correlation (Corr) scores on both datasets, affirming its adeptness in discerning sentiment-pertinent features. This advancement not only highlights the model's analytical precision but also solidifies its reputation for delivering reliable and robust performance.

One of the most remarkable achievements of InfoEnh is its exemplary performance in the realm of fine-grained sentiment classification, represented as "Acc-7". This task, recognized for its complexity, sees InfoEnh significantly surpassing the baseline, as evidenced by a remarkable 1.18% improvement in ConFEDE on the MOSEI dataset. This notable enhancement can be ascribed to the model's advanced feature filtering and alignment strategies, meticulously designed to identify and accentuate intricate sentiment nuances. In summary, the results provide strong evidence of InfoEnh's consistent capacity to enhance the performance.

5.3. Ablation Study

5.3.1. Effects of Different Components

To assess the impact of individual components within InfoEnh, a series of ablation experiments on MOSEI and MOSEI was conducted, and the results are summarized in Table 2. We selected self-MM and ConFEDE for comparison.

An initial observation from our experiments highlighted a notable decrease in performance upon the removal of FIB. Specifically, when we remove the FIB and directly utilize the top-k embeddings, the model's effectiveness markedly diminished. Furthermore, we explored the impact of removing the FIA, which is instrumental in aligning features across different modalities. These experiments consistently demonstrated the effectiveness of using optimal transport to align the features on different domains.

The ablation results unequivocally demonstrate a performance decrement with the removal of each component, particularly with FIB, which highlights its critical role in the model's success. These findings robustly validate the importance of the individual elements incorporated into InfoEnh, offering clear evidence of their substantial contributions to the model's superior performance and affirming the model's overall efficacy and resilience in multimodal sentiment analysis tasks.

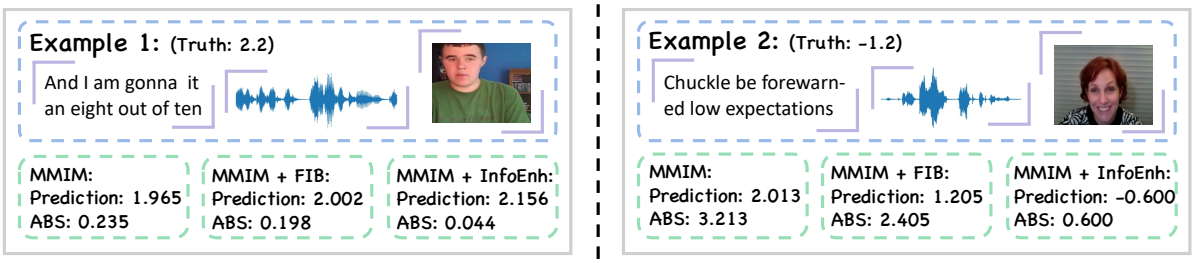

Figure 5: Case study of the baseline added InfoEnh's component.

5.3.2. Effects of Different Modalities

Table 3 showcases the results of an ablation study designed to assess the contribution of each individual modality, aiming to elucidate their respective impacts more thoroughly. We methodically removed one modality from the baseline model. The findings consistently indicated that the omission of any modality, particularly the text modality, led to a significant decline in performance across various evaluation metrics. This observation underscores the intrinsic value of a multimodal approach, where each modality contributes distinctively to the analysis, thereby emphasizing the collaborative effect of diverse input data.

Furthermore, the integration of our InfoEnh component into the baseline model demonstrated its capability to mitigate the effects of missing modalities. Notably, in scenarios where the critical text modality was absent, the incorporation of InfoEnh resulted in substantial enhancements, with improvements of 11.52% in Acc-2 and 10.35% in F1 scores. This underscores InfoEnh's effectiveness and robustness in bolstering the model's resilience against the absence of key modalities.

5.4. Case Study

To further illustrate the advantages of our model compared to existing baselines in MSA, we present case studies as depicted in Figure 5. We initially integrate the FIA module, followed by the comprehensive integration of the InfoEnh module. These case studies reveals that the previous model occasionally incurs significant errors, which are markedly reduced after the stepwise addition of InfoEnh components. These improvements underscore our proposed InfoEnh's capability in enhancing the accuracy of sentiment predictions, showcasing its utility in refining and optimizing sentiment analysis processes.

5.5. Interpretability of InfoEnh

5.5.1. Visualization of Heat Map

In this section, we delve into the interpretability of InfoEnh by employing Grad-CAM (Selvaraju et al., 2017) to generate heatmaps for images and text. As shown in Figure 6, within the image modality, the smile of the little boy is assigned the highest weight, signifying its importance. Similarly, in the text modality, the term "great", with its positive connotation, receives the highest weight. These key elements adeptly encapsulate the sentiment inherent in each modality, providing a clear and concise representation of the underlying emotional content. The precision and focus exhibited in these heatmaps can be largely attributed to the functionality of the FIA module within InfoEnh. This component plays a critical role in sifting through the modalities to discard any irrelevant or superfluous information, thereby highlighting the features that are most salient and consequential.

Figure 6: Visual explanations and textual attention map for a sample from MOSI dataset.

5.5.2. Visualization of Modality Representations

In Figure 7, we present a visualization of the modal embeddings within a 2D feature space, utilizing the t-SNE technique (van der Maaten and Hinton, 2008) on the MOSI dataset. We compare the embeddings before and after the integration of InfoEnh into the baseline. Following the integration of InfoEnh, it is evident that the inter-modality information becomes more cohesive, with a notable reduction in the distance between modalities, illustrating the InfoEnh's effectiveness in fostering closer alignment and concentration of modal information. It proves that InfoEnh plays a pivotal role in refining and elevating the quality of MSA by en-

suring a more harmonious and interconnected representation of data across varying modalities.

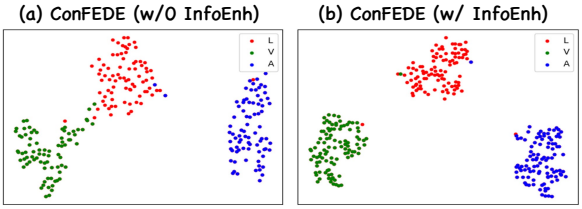

Figure 7: Visualization of modality representations in 2D space by using t-SNE.

6. Conclusion

The introduced InfoEnh in this work addresses significant challenges in multimodal sentiment analysis by effectively filtering irrelevant data and aligning features across different modalities using a novel approach. The proposed method demonstrates improved performance on benchmark datasets, showcasing its potential to enhance the accuracy and robustness of sentiment analysis models. The integration of feature information bottleneck and optimal transport alignment contributes to the model's ability to capture nuanced sentiment cues and maintain consistency across modalities. The experimental results and ablation studies validate the importance of each component, highlighting the overall efficacy of InfoEnh in multimodal sentiment analysis.

7. References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *Proc. of ICLR*.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Proc. of EMNLP (Findings)*.

Gary Ang and Ee-Peng Lim. 2022. Guided attention multimodal multitask financial forecasting with inter-company relationships and global and local news. In *Proc. of ACL*.

Rishabh Bhardwaj, Tushar Vaidya, and Soujanya Poria. 2022. KNOT: knowledge distillation using optimal transport for solving NLP tasks. In *Proc. of COLING*.

Jie Cao and Yin Zhang. 2022. Otseq2set: An optimal transport enhanced sequence-to-set model for extreme multi-label text classification. In *Proc. of EMNLP*.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: multi-modal knowledge graph embeddings via optimal transport. In *Proc. of NeurIPS*.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proc. of ACL*.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2023. PLOT: prompt learning with optimal transport for vision-language models. In *Proc. of ICLR*.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019a. Improving sequence-to-sequence learning via optimal transport. In *Proc. of ICLR*.

Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019b. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proc. of SIGIR*.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, Xianwei Zhuang, and Yuexian Zou. 2024. Towards multi-intent spoken language understanding via hierarchical attention and optimal transport. In *Proc. of AAAI*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. of NeurIPS*.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *Proc. of FUSION*.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proc. of ACM MM*.

Jun Han and Claudio Moraga. 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *Proc. of IWANN*.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proc. of EMNLP*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *Proc. of ACM MM*.

Zhiqi Huang, Dongsheng Chen, Zhihong Zhu, and Yuexian Zou. 2023. Mclf: A multi-grained contrastive learning framework for asr-robust spoken language understanding. In *Proc. of EMNLP (Findings)*.

Leonid V Kantorovich. 2006. On the translocation of masses. *Journal of mathematical sciences*.

Huisheng Mao, Baozheng Zhang, Hua Xu, Ziqi Yuan, and Yihe Liu. 2023. Robust-msa: Understanding the impact of modality noise on multimodal sentiment analysis. In *Proc. of AAAI*.

Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrusaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proc. of ICMI*.

Nasim Nouri. 2022. Text style transfer via optimal transport. In *Proc. of NAACL*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *Proc. of NeurIPS*.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Proc. of ICDM*.

Shraman Pramanick, Aniket Roy, and Vishal M. Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proc. of WACV*.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Proc. of ACL (Findings)*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of ICCV*.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv*.

Cees Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proc. of ACM MM*.

Florian Spiess and Heiko Schuldt. 2022. Multimodal interactive lifelog retrieval with vitrivr-vr. In *Proc. of LSC@ICMR*.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *ITW*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*.

Leonid Nisonovich Vaserstein. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*.

Cédric Villani. 2021. *Topics in optimal transportation*. American Mathematical Soc.

Binqiang Wang, Gang Dong, Yaqian Zhao, Rengang Li, Qichun Cao, and Yinyin Chao. 2022. Non-uniform attention network for multi-modal sentiment analysis. In *Proc. of MMM*.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *Proc. of ICME*.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proc. of ACL*.

Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising bottleneck with mutual information maximization for video multimodal fusion. In *Proc. of ACL*.

Yifeng Xie, Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, and Dongsheng Chen. 2023. Syntax matters: Towards spoken language understanding via syntax-aware attention. In *Proc. of EMNLP (Findings)*.

Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and Philip S. Yu. 2021a. Social image sentiment analysis by exploiting multimodal content and heterogeneous relations. *IEEE Transactions on Industrial Informatics*.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021b. Vocabulary learning via optimal transport for neural machine translation. In *Proc. of ACL/IJCNLP*.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2023. A deep multi-level attentive network for multimodal sentiment analysis. *ACM Trans. Multim. Comput. Commun. Appl.*

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proc. of ACM MM*.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proc. of ACL*.

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 *Proc. of AAAI*.

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 *Proc. of ACL (Findings)*.

Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, and Mei Chen. 2024. Rethinking multimodal content moderation from an asymmetric angle with mixed-modality. In *Proc. of WACV*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proc. of EMNLP*.

Amir 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 *Proc. of ACL*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.*

Yufei Zeng, Zhixin Li, Zhenjun Tang, Zhenbin Chen, and Huifang Ma. 2023. Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. *Expert Syst. Appl.*

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 *Proc. of EMNLP*.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Enhancing code-switching for cross-lingual slu: A unified view of semantic and grammatical coherence. In *Proc. of EMNLP*.

Zhihong Zhu, Yunyan Zhang, Xuxin Cheng, Zhiqi Huang, Derong Xu, Xian Wu, and Yefeng Zheng. 2024. Alignment before awareness: Towards visual question localized-answering in robotic surgery via optimal transport and answer semantics. In *Proc. of COLING*.