FINE: Factorized Multimodal Sentiment Analysis via Mutual INformation Estimation

Yadong Liu, Shangfei Wang*

University of Science and Technology of China
yadongliu@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract

Multimodal sentiment analysis remains a challenging task due to the inherent heterogeneity across modalities. Such heterogeneity often manifests as asynchronous signals, imbalanced information between modalities, and interference from task-irrelevant noise, hindering the learning of robust and accurate sentiment representations. To address these issues, we propose a factorized multimodal fusion framework that first disentangles each modality into shared and unique representations, and then suppresses task-irrelevant noise within both to retain only sentiment-critical representations. This fine-grained decomposition improves representation quality by reducing redundancy, prompting cross-modal complementarity, and isolating task-relevant sentiment cues. Rather than manipulating the feature space directly, we adopt a mutual information–based optimization strategy to guide the factorization process in a more stable and principled manner. To further support feature extraction and long-term temporal modeling, we introduce two auxiliary modules: a Mixture of Q-Formers, placed before factorization, which precedes the factorization and uses learnable queries to extract fine-grained affective features from multiple modalities, and a Dynamic Contrastive Queue, placed after factorization, which stores latest high-level representations for contrastive learning, enabling the model to capture long-range discriminative patterns and improve class-level separability. Extensive experiments on multiple public datasets demonstrate that our method consistently outperforms existing approaches, validating the effecti veness and robustness of the proposed framework.

Introduction

Sentiment analysis aims to uncover sentiments or opinions when individuals encounter specific topics, people, or entities (Soleymani et al. 2017). Since its inception, it has rapidly evolved into a vital research area with widespread applications in robotics, healthcare, education, and other industries (Melville, Gryc, and Lawrence 2009; Petrovica, Anohina-Naumeca, and Ekenel 2017; Liu et al. 2017; Sánchez-Rada and Iglesias 2019). As the internet transitions from text-based to a multimedia-driven space, sentiment analysis has dramatically transformed, leading to the rise of multimodal sentiment analysis (MSA). MSA integrates

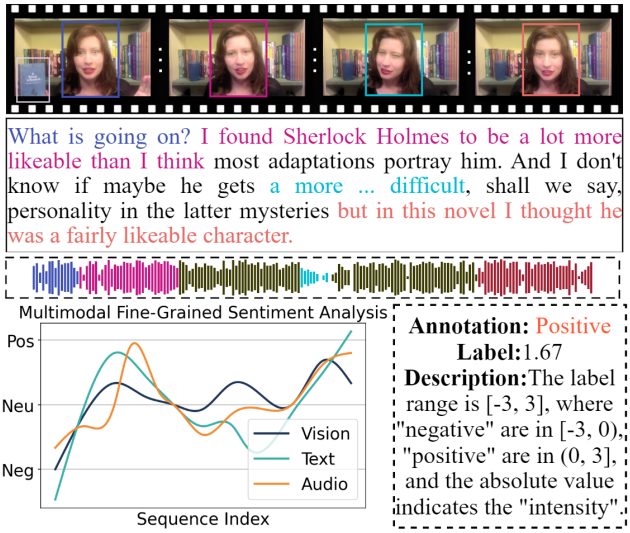


Figure 1: A sample of MSA, incorporating three modalities: visual, textual, and audio. The bottom-left section displays fine-grained sentiment analysis, while the bottom-right section shows the label and annotation for this example.

rich information from diverse modalities like text, audio, and images, providing a more comprehensive and accurate understanding of human sentiment (Gandhi et al. 2023). However, significant heterogeneity exists across modalities. Each modality not only reveals distinct types of sentiment cues but also varies in representational density and noise. For example, text typically provides high semantic density (He et al. 2022), while video and audio may contain redundant or overly complex sentiment representations.

MSA involves developing fusion strategies across different modalities to comprehensively predict sentiment polarity and intensity (Morency, Mihalcea, and Doshi 2011). Recent works primarily focus on multimodal representation learning and fusion, aiming to encode and integrate representations from diverse modalities to identify sentiment patterns. To achieve effective prediction, a range of approaches have been explored, including Multi-Layer Perceptrons (MLPs), attention-based models (Sun et al. 2022), Long Short-Term Memory (LSTM) networks (Tsai et al. 2018; Lin and Hu

*Corresponding author.

2024), Graph Neural Networks (Mai, Hu, and Xing 2020), and transformer encoders (Tsai et al. 2019; Rahman et al. 2020; Han et al. 2021; Liang et al. 2022). Among these techniques, fusion strategies play a central role in aligning features from different modalities.

From the perspective of fusion strategies, MSA methods are commonly categorized into early, intermediate, and late fusion. Early fusion (Liu et al. 2018; Poria et al. 2016) directly concatenates features from different modalities, but often fails to capture inter-modal heterogeneity. Late fusion (Zhang et al. 2023) combines predictions from unimodal classifiers, yet lacks inter-modal interactions. Intermediate fusion (Mai, Hu, and Xing 2020; Nagrani et al. 2021; Mai, Zeng, and Hu 2022; Ma, Zhang, and Sun 2023; Fan et al. 2023; Jiang et al. 2024) offers a flexible trade-off, enabling joint embedding learning while reducing noise. Notably, models like Tensor Fusion Networks (Zadeh et al. 2017) and Multimodal Transformers (Tsai et al. 2019) have demonstrated success in joint modeling. However, many of these methods treat each modality as a whole and perform fusion through simple operations such as concatenation or weighted averaging in a shared latent space. Such strategies often fail to capture modality-specific information and limit the model's flexibility and interpretability in handling modality heterogeneity.

To more effectively address this issue, feature disentanglement has gained popularity. Representative works such as MISA (2020) project modality representations into two distinct subspaces. The first subspace captures modality-invariant features by aligning common semantics, while the second preserves modality-specific representations that reflect the private characteristics of each modality. DMD (2023) introduces a graph distillation unit to dynamically decouple homogeneous and heterogeneous features. ConFEDE (Yang et al. 2023) further decomposes modality features into similar and dissimilar features and employs contrastive loss to unify representation learning with disentanglement. However, these methods have yet to fully exploit the potential of disentanglement, as two subspaces may still contain task-irrelevant noise. Directly using them may result in optimization conflicts between the reconstruction and classification tasks. To address this, we design a two-stage disentanglement module. The first stage follows existing formulations by factorizing each modality into shared and unique representations and performing reconstruction. The second stage suppresses task-irrelevant noise in both branches, retaining only task-relevant representations for downstream sentiment modeling. Unlike previous approaches that directly manipulate the feature space using metrics such as Central Moment Discrepancy, Frobenius norm minimization, or contrastive loss, our method adopts a mutual information estimation objective, providing a more stable foundation for disentanglement.

As illustrated in Figure 1, multimodal sentiment cues are often temporally heterogeneous. In the visualization, we align content from different modalities using consistent color coding—segments with the same color correspond to the same time step. At the beginning, the segment *"What is going on?"* expresses mild confusion in both audio and text, but it deviates from the final sentiment label and acts as task-irrelevant noise. The phrase *"I found Sherlock Holmes to be a lot more likeable..."* conveys consistent positivity across all three modalities, reflecting shared task-relevant cues. However, the facial expression and intonation corresponding to this sentiment vary in timing and intensity, indicating asynchronous signals and imbalanced information across modalities. Later, in *"a more... difficult,"* sentiment becomes ambiguous: the text implies negativity, while the speaker's playful expression and tone suggest otherwise. This highlights inconsistency in sentiment across modalities and suggests the presence of unique information in each channel. Only by integrating all modalities can this subtle positivity be accurately interpreted, emphasizing the importance of cross-modal complementarity. These observations suggest that multimodal sentiment analysis faces three key challenges: asynchronous signals, imbalanced information, and interference from task-irrelevant noise. These issues stem from inherent modality heterogeneity, which hinders the learning of robust and accurate sentiment representations. Therefore, there is a critical need for a principled fusion strategy that can disentangle shared and unique features, suppress task-irrelevant noise, and leverage cross-modal synergy to reduce redundancy, enhance complementarity, and improve sentiment modeling.

To overcome these issues, we propose FINE, a factorized multimodal sentiment analysis framework grounded in mutual information estimation. FINE first employs a Mixture of Q-Formers to extract fine-grained sentiment cues from each modality at an early stage using learnable queries. These extracted features are then fed into the Factorized Task-Relevant Encoder, which factorizes each extracted feature into shared and unique representations, and further removes task-irrelevant noise from both by optimizing mutual information objectives. A Transformer encoder is subsequently used to fuse the task-relevant features across modalities, integrating semantic and affective cues. To support long-term pattern modeling and improve representation robustness, we introduce a Dynamic Contrastive Queue that stores recent high-level features for temporal contrastive learning. Together, these components enable FINE to effectively model complex, noisy, and asynchronous multimodal sentiment signals. Our contributions are summarized as follows:

- We propose a factorized framework that jointly disentangles shared/unique and task-relevant/irrelevant features, enhancing robustness and interpretability.
- We introduce a query-based extraction mechanism for fine-grained sentiment representations and a contrastive learning module for capturing long-range dependencies.
- Experiments on multiple benchmarks validate the effectiveness of FINE and demonstrate the utility of each component.

Methodology

As illustrated in Figure 2, FINE first processes raw modality features using a set of Q-Former experts (MoQ). Each expert employs L learnable query tokens to extract temporally

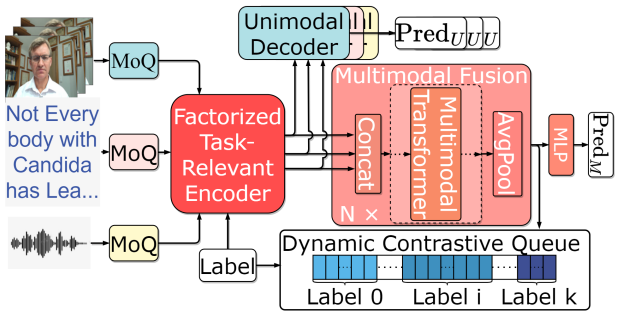


Figure 2: Overview of FINE. Pred_U and Pred_M denote the unimodal and multimodal predictions, respectively.

aligned, fine-grained sentiment cues from the original multimodal sequences, effectively mitigating cross-modal temporal inconsistencies. The weighted query results from each modality are then passed to the Factorized Task-Relevant Encoder, which disentangles them into shared and unique task-relevant features. These six types of features (two from each modality) are concatenated and fed into a Transformer encoder to capture high-level semantic and sentiment dependencies across modalities. Finally, FINE introduces a Dynamic Contrastive Queue, which stores the latest task-relevant representations in class-aware sub-queues, promoting better discrimination across classes and preserving meaningful distance structures that reflect class-level distinctions. The following subsections detail the structure and functionality of each module.

Mixture of Q-Formers

To effectively capture modality-specific and fine-grained sentiment cues, we integrate a Mixture-of-Experts architecture into our framework. MoQ is a straightforward extension of MoE for multimodal sentiment analysis, where each expert is designed to model diverse input subspaces across modalities. Instead of using conventional feedforward networks (FFNs) as experts, we employ Q-Formers (Li et al. 2023), which act as information bottlenecks by leveraging learnable query tokens to extract condensed sentiment representations. This design enables MoQ to retain expressive capacity for nuanced sentiment features while significantly reducing computational overhead. To ensure routing flexibility and prevent the model from collapsing into static expert usage, we adopt an auxiliary load-balancing loss $\mathcal{L}_{aux}$, which encourages more uniform expert activation during training.

Formally, for a modality feature $X_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{T, A, V\}$ denotes text, audio, or visual modality, the MoQ module produces a compressed output of l tokens:

$$\hat{x}_m^i = \text{MoQ}_m(x_m^i), \tag{1}$$

where $\hat{x}_m^i \in \mathbb{R}^{l \times d'_m}$, and d'_m denotes the hidden dimensionality of the Q-Former outputs.

Further details on expert routing and the internal structure of Q-Formers are provided in the Supplementary Material.

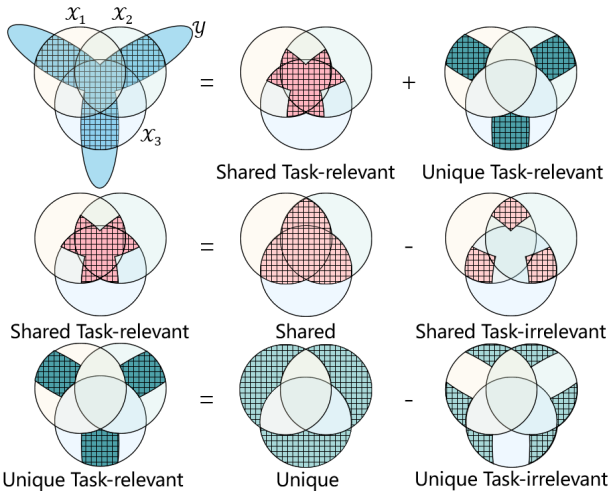


Figure 3: A Venn diagram illustrating mutual information among modalities $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ and the task $\mathcal{Y}$. The blue region denotes task-relevant information, with the blue grid marking the ideal target for learning. This target combines task-relevant shared information (red) and unique information (green). Achieving it requires suppressing task-irrelevant noise from both components.

Factorized Task-Relevant Encoder

Factorized Task-Relevant Encoder (FTRE) is designed as a general-purpose module applicable to a broad range of tri-modal supervised tasks in real-world scenarios. Given an input space consisting of three modalities X_1, X_2, and X_3, FTRE assumes that the task-relevant information can be factorized into two types: shared information and unique information. The former denotes information common across multiple modalities, while the latter captures information specific to individual modalities. Both types of information are essential for accurately modeling the target variable Y.

As demonstrated in the Venn diagram in Figure 3, the ideal task-relevant information is factorized into two conditional mutual information terms in the tri-modal feature space: one representing the shared task-relevant information S_{TR}, and the other representing the unique task-relevant information U_{TR}. This factorization is achieved in two steps, as follows:

$$I(X_1, X_2, X_3; Y) = S_{TR} + U_{TR} = (S - S_{TI}) + (U - U_{TI}), \tag{2}$$

where S represents the shared information between any two modalities X_i and X_j, and U represents the unique information specific to each modality. Specifically, S can be expressed as the aggregation of pairwise mutual information across modalities. On the other hand, U can be expressed as the sum of the conditional mutual information between one modality and the rest:

$$S = \sum_{1 \le i < j \le 3} I(X_i; X_j), U = \sum_{\substack{\{i,j,k\} \subseteq \{1,2,3\} \\ i \ne j \ne k}} I(X_i \mid X_j, X_k), \tag{3}$$

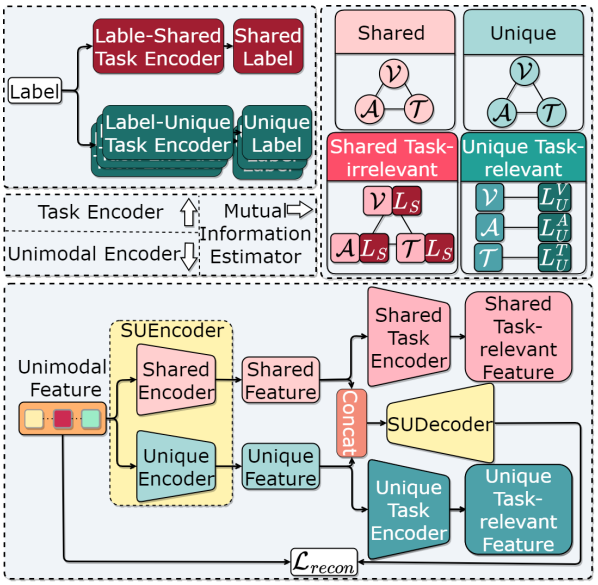


Figure 4: The structure of FTRE. The bottom-left corner depicts the unimodal encoders, where each modality X_i is processed independently. The top-left corner illustrates the direct encoding of labels, resulting in one shared and three unique label embeddings. The top-right corner represents the estimation of four distinct types of mutual information.

and S_{TR} represents the mutual information corresponding to the shared task-relevant information. In contrast, S_{TI}, which is irrelevant to the task, represents noisy information that should be excluded. S_{TI} can be defined as the union of pairwise mutual information conditioned on the task Y. Directly solving for S_{TR} is not very practical, so we indirectly approximate the S_{TR} by minimizing S_{TI}:

$$S_{TI} = \sum_{1 \le i < j \le 3} I(X_i; X_j \mid Y)\,, \tag{4}$$

and U_{TR} denotes the unique task-relevant information. This term highlights the distinct contributions of each modality to the task and can be expressed as the union of the mutual information between each modality and the task:

$$U_{TR} = \sum_{1 \le i \le 3} I(X_i; Y)\,. \tag{5}$$

We now move from the above theoretical analysis to the practical implementation of FTRE for extracting S_{TR} and U_{TR}. The detailed architecture of FTRE is shown in Figure 4. Within the Unimodal Encoder, the proposed Shared&Unique Encoder (SUEncoder) projects input modality features X_m into shared S_m and unique U_m representations. The SUEncoder comprises MLPs with activation functions, namely the shared encoder (E^s_m) and the unique encoder (E^u_m).

$$x^s_{m,i} = E^s_m(\hat{x}^i_m)\,, \quad x^u_{m,i} = E^u_m(\hat{x}^i_m)\,. \tag{6}$$

To compute S in Equation (2), we introduce the InfoNCE (Oord, Li, and Vinyals 2018) to maximize a lower bound on the mutual information between shared features $x^s_{m,i}$ from any two modalities,

$$I_{sha}\left(X^s_1; X^s_2\right) = \mathbb{E}_{\substack{x_1, x_2^+ \sim p(x_1,x_2) \\ x_2^- \sim p(x_2)}} \left[\log \frac{\exp f\left(x_1, x_2^+\right)}{\sum_k \exp f(x_1, x_2^-)}\right], \tag{7}$$

Where $f(x_1, x_2^+)$ is the optimal critic, and x_2^+ refers to the feature of another modality from the same sample as x_1, while x_2^- denotes the feature from a different sample.

To compute U in Equation 2, the NCE-CLUB (Liang et al. 2023) is introduced, which minimizes the expected upper bound on the unique feature $x^u_{m,i}$ between any two modalities. It effectively achieves this "for free" while avoiding the need to separately optimize both the lower bound and upper bound:

$$\begin{aligned} I_{uni}(X^u_1; X^u_2) = {} & \mathbb{E}_{x_1, x_2^+ \sim p(x_1,x_2)} \left[f^*(x_1, x_2^+)\right] \\ & - \mathbb{E}_{\substack{x_1 \sim p(x_1) \\ x_2^- \sim p(x_2)}} \left[f^*(x_1, x_2^-)\right], \end{aligned} \tag{8}$$

where $f^*(x_1, x_2^+)$ is the optimal critic from I_{NCE}, used within the I_{CLUB} (Cheng et al. 2020).

The aforementioned operations establish a disentanglement mechanism, ensuring that both representations retain as much of information from the original modality as possible. To reinforce information preservation, we incorporate a decoder that reconstructs the original modality features $\hat{X}_m$ from the representations of the shared and unique information. The reconstruction process is guided by a reconstruction loss $\mathcal{L}_{recon}$, calculated using Mean Squared Error, which encourages accurate recovery of the modality-specific input. To further suppress task-irrelevant noise, we feed the unique and shared features, $x^u_{m,i}$ and $x^s_{m,i}$, into their respective task encoders, E^{utr}_m and E^{str}_m. Both E^{utr}_m and E^{str}_m are implemented as MLPs, which serve to refine the representations by filtering out noise while preserving task-relevant semantics, as formulated in the following equations:

$$x^{str}_{m,i} = E^{str}_m(x^s_{m,i})\,, \quad x^{utr}_{m,i} = E^{utr}_m(x^u_{m,i})\,. \tag{9}$$

To ensure the effective realization of the aforementioned functionality, we introduce the conditional InfoNCE estimator and the NCE-CLUB estimator as constraints to guarantee that the extracted information remains task-relevant. Specifically, for the concatenated feature set of $X^{str}_m \in \mathbb{R}^d$ from modality m and Y^{str}, where $Y^{str} \in \mathbb{R}^{d_s}$ is obtained by passing the original label Y through a Label-Shared Task Encoder, the mutual information between this feature set and the concatenated feature set of X^{str}_n from another modality n along with Y^{str} is minimized in its upper bound using NCE-CLUB. This procedure effectively minimizes the task-irrelevant mutual information:

$$\begin{aligned} I_{str}(X^{str}_1; X^{str}_2 \mid Y^{str}) = \mathbb{E}_{p(y)} \Big[& \\ & \mathbb{E}_{x_1, x_2^+ \sim p(x_1,x_2|y)} \left[f^*(x_1, x_2^+, y)\right] \\ & - \mathbb{E}_{\substack{x_1 \sim p(x_1|y) \\ x_2^- \sim p(x_2|y)}} \left[f^*(x_1, x_2^-, y)\right] \Big]. \end{aligned} \tag{10}$$

Similarly, we employ the InfoNCE estimator to maximize the lower bound of the mutual information between $X_m^{utr} \in \mathbb{R}^d$ from modality m and $Y_m^{utr} \in \mathbb{R}^{d_u}$, where Y_m^{utr} is obtained by passing the original label Y through a Label-Unique Task Encoder. This operation serves to reduce the task-irrelevant noise present within the unique features:

$$I_{utr}\left(X_1^{utr}; Y_1^{utr}\right) = \mathbb{E}_{\substack{x_1, y_1^+ \sim p(x_1, y_1) \\ y_1^- \sim p(y_1)}} \left[\log \frac{\exp f\left(x_1, y_1^+\right)}{\sum_k \exp f(x_1, y_1^-)} \right], \tag{11}$$

where y_1^+ represents a positive label related to x_1, while y_1^- is a negative label from the marginal distribution $p(y)$.

The total mutual information loss integrates the four mutual information estimations discussed above. Specifically, I_{sha} and I_{utr} are estimated as lower bounds and optimized by minimizing their negatives, while I_{uni} and I_{str} are treated as upper bounds and directly minimized. The overall mutual information loss $\mathcal{L}_{MI}$ is given by:

$$\mathcal{L}_{MI} = -I_{sha} + I_{uni} + \mathcal{L}_{recon} + I_{str} - I_{utr}. \tag{12}$$

The final output of each modality is obtained by concatenating X_m^{str} and X_m^{utr}, thereby integrating task-relevant information extracted from both shared and unique features.

Dynamic Contrastive Queue

To capture long-term discriminative patterns and mitigate incidental noise, we incorporate a Dynamic Contrastive Queue strategy. This queue-based mechanism uses class-wise subqueues to retain recent task-relevant features over time. It allows the model to encourage greater inter-class representation separation, effectively addressing class imbalance and noise sensitivity. Additionally, we adopt the Angle-Compensated Contrastive Regularizer (ACCon) (Zhao et al. 2025) to adjust similarity based on label differences, improving representation precision in continuous sentiment spaces. Finally, at training step t, we compute the angle-compensated contrastive loss $\mathcal{L}_{CL}^i$ for sample i using the updated queue:

$$\mathcal{L}_{CL}^i = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(z_i z_p^T / \tau)}{\left(\begin{array}{c} \sum_{k \in \mathcal{P}(i)} \exp(z_i z_k^T / \tau) \\ + \sum_{m \in \mathcal{N}(i)} \exp(\cos(\tilde{\theta}_{i,m}) / \tau) \end{array} \right)}, \tag{13}$$

where $\mathcal{P}(i)$ and $\mathcal{N}(i)$ represent the positive and negative sample sets for anchor i, sampled from both the current batch at step t and the historical queue accumulated over the previous $t-1$ steps. Further details, including the queue update process and angle-aware similarity compensation, are provided in the Supplementary Material.

Multimodal Fusion and Prediction

We combine the shared task-relevant features and the unique task-relevant features to construct multimodal representations for downstream tasks, which include both auxiliary unimodal prediction and multimodal fusion-based prediction. To extract and fuse these multimodal features, we leverage the Transformer (Vaswani et al. 2017) to learn comprehensive multimodal sentiment knowledge. Additionally, we use the [CLS] token of each each modality as a representation of the information of that modality. Then, fully connected layers are applied to predict the final results based on these fused multimodal representations.

We integrate the above losses to formulate the comprehensive optimization objective:

$$\begin{aligned} \mathcal{L}_{total} = & \mathcal{L}_{MP} + \lambda_{up}\mathcal{L}_{UP} + \lambda_{cl}\mathcal{L}_{CL} \\ & + \lambda_{aux}\mathcal{L}_{aux} + \beta_{mi}\mathcal{L}_{MI}, \end{aligned} \tag{14}$$

where $\mathcal{L}_{MP}$ and $\mathcal{L}_{UP}$ represent the multimodal and unimodal prediction losses, respectively, defined as the Mean Squared Error in our experiments. The hyperparameters λ_{cl}, λ_{up}, λ_{aux}, and β_{mi} control the relative contribution of the different loss components to the overall optimization.

Experiments

Experimental Settings

Dataset To validate the effectiveness of the proposed model, we conducted experiments on four widely-used multimodal datasets: CMU-MOSI (2016), CMU-MOSEI (2018), UR-FUNNY (Hasan et al. 2019), and CH-SIMS (Yu et al. 2020). These datasets span a variety of domains and languages, enabling a comprehensive evaluation of model performance under different multimodal sentiment analysis scenarios. Detailed descriptions of the datasets, evaluation metrics, implementation settings, and baseline models can be found in the supplementary material.

Results and Analysis

Quantitative Results Table 2 and Table 1 present a comparative analysis of FINE with recent state-of-the-art models on the CMU-MOSEI and CMU-MOSI datasets. The best results are highlighted in bold font, and the second-best results are underlined. As shown in the table, on CMU-MOSEI, FINE yields consistent gains across all metrics, with 1-point improvements on ACC-2 and F1, and a 0.7-percentage-point gain on ACC-7. On CMU-MOSI, FINE sets new benchmarks in ACC-2, F1, and ACC-7. We hypothesize that the limited size of the CMU-MOSI dataset

Model	ACC-2 (↑)	F1 (↑)	ACC-7 (↑)
TFN (Zadeh et al. 2017)	80.8	80.7	34.9
LMF (Liu et al. 2018)	82.5	82.4	33.2
GFN (2020)	84.3	84.3	47.0
MulT (Tsai et al. 2019)	83.7	83.7	41.5
CubeMLP (Sun et al. 2022)	85.6	85.5	45.5
MISA	83.4	83.6	42.3
BBFN (Han et al. 2021)	84.3	84.3	45.0
C-MIB (2022)	85.2	85.2	<u>48.2</u>
MSG (Lin and Hu 2024)	85.7	85.6	45.3
ConFEDE (Yang et al. 2023)	85.52	85.52	42.27
DMD (2023)	86.0	86.0	45.6
EUAR (Gao et al. 2024)	<u>86.3</u>	<u>86.3</u>	46.1
FINE (Ours)	**86.95**	**86.94**	**48.50**

Table 1: Results on CMU-MOSI dataset.

Model	ACC-2 (↑)	F1 (↑)	ACC-7 (↑)
TFN (Zadeh et al. 2017)	82.5	82.1	50.2
LMF (Liu et al. 2018)	82.0	82.1	48.0
GFN (2020)	85.0	85.0	51.8
MulT (Tsai et al. 2019)	84.7	84.6	50.7
CubeMLP (Sun et al. 2022)	85.1	84.5	54.9
MISA (2020)	85.5	85.3	52.2
BBFN (Han et al. 2021)	86.2	86.1	54.8
C-MIB (2022)	86.2	86.2	53.0
MSG (Lin and Hu 2024)	85.4	85.4	52.8
ConFEDE (Yang et al. 2023)	85.82	85.83	54.86
DMD (2023)	86.6	86.6	54.5
EUAR (Gao et al. 2024)	86.6	86.4	54.9
FINE (Ours)	**87.70**	**87.68**	**55.59**

Table 2: Results on CMU-MOSEI datasets.

Model	Context	Target	ACC-2 (↑)
C-MFN (Hasan et al. 2019)	✓	✓	65.23
LMF (Liu et al. 2018)		✓	67.53
TFN (Zadeh et al. 2017)		✓	68.57
MISA (2020)		✓	70.61
MAGBERT(XLNet) (2020)		✓	72.43
MuLOT (2022)		✓	73.22
MuLOT (2022)	✓	✓	73.97
FINE (Ours)		✓	**74.95**

Table 3: Results on UR-FUNNY datasets.

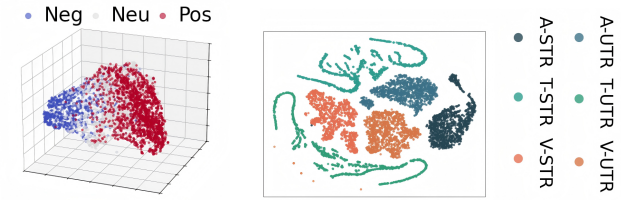
may have limited the expressiveness of FINE. Table 3 evaluates FINE on the multimodal humor recognition dataset UR-FUNNY. FINE achieves a new state-of-the-art ACC-2 score of 74.95%, outperforming the best previous model MAGBERT by 2.5 points. In Table 4, we also report results on CH-SIMS, a challenging Chinese-language sentiment benchmark with real-world video data. FINE achieves the best results on both ACC-2 and F1, and ranks second on ACC-5. These results demonstrate the effectiveness of our method in cross-lingual and more diverse multimodal settings. These improvements can be attributed to the disentanglement mechanism in FINE, which explicitly separates task-relevant shared and unique information while suppressing noise. Furthermore, MoQ facilitates early-stage alignment of multimodal representations, and DCQ improves long-range dependency modeling and class-level discrimination. Together, these design choices allow FINE to learn robust and complementary representations, leading to its superior performance across diverse benchmark settings.

Visualization of Representations Additionally, we visualized the representations obtained from the FTRE module, which fuses all modalities. Each modality's features were further decomposed into Shared Task-Relevant (STR) and Unique Task-Relevant (UTR) features. These six sets of features were concatenated and then used for classification visualization. The results are shown in Figure 5, based on the test set of the CMU-MOSEI dataset. In Figure 5(a), we present the 3D visualization between learned features and

Model	ACC-2 (↑)	F1 (↑)	ACC-5 (↑)
LF-DNN (Yu et al. 2020)	78.87	79.87	41.62
MFN(A) (Zadeh et al. 2018)	78.87	79.87	39.47
LMF (Liu et al. 2018)	77.77	77.88	40.53
TFN (Zadeh et al. 2017)	78.38	78.62	39.30
MulT(A) (Tsai et al. 2019)	78.38	78.62	37.94
Self-MM (Yu et al. 2021)	80.04	80.44	41.53
ConFEDE (Yang et al. 2023)	82.23	82.08	**46.30**
Coupled Mamba (2024)	81.8	81.3	42.1
FINE (Ours)	**82.28**	**82.22**	41.79

Table 4: Results on CH-SIMS datasets. (A) means the model utilized the aligned data.

sentiment labels. Blue, red, and gray represent Negative, Positive, and Neutral classes, respectively. From the density of color distributions and their spatial layout, it is evident that the representations learned by the FTRE module are strongly correlated with sentiment polarity. Moreover, Neutral points are mostly located between the Negative and Positive clusters, which aligns well with human sentiment intuition.



(a) Classification (b) Different Modalities

Figure 5: Visualization of features obtained after FTRE.

Figure 5(b) further illustrates that the six types of features exhibit both strong separability and cross-modal complementarity. Notably, STR and UTR features from the same modality display similar distributions, while those from different modalities are distinctly separated. This suggests that our method preserves the structural characteristics of each modality without enforcing strong constraints in the feature space. Furthermore, all six feature types show relatively high internal density, indicating that they capture high-level sentiment semantics while effectively filtering out irrelevant noise.

Ablation Study

In this section, we analyze the impact of each modality on performance using the CMU-MOSEI dataset and conduct ablation studies on key components of FINE. To assess their importance, we replace each with functionally similar alternatives and report results in Table 5 and Table 6.

Impact of Different Modalities Table 5 reports the performance of different modality combinations on CMU-MOSEI. Results show that adding more modalities consistently improves performance, with all three modalities combined yielding the best results. Among them, text contributes

Modality			Metrics		
V	A	T	ACC-2 (↑)	F1 (↑)	ACC-7 (↑)
✓	✗	✗	63.32	61.47	41.38
✗	✓	✗	63.35	58.48	39.51
✗	✗	✓	85.14	85.32	53.40
✓	✓	✗	65.24	63.13	39.60
✓	✗	✓	86.24	86.21	53.42
✗	✓	✓	86.98	86.78	54.13
✓	✓	✓	**87.70**	**87.68**	**55.59**

Table 5: Ablation studies for different multimodal fusion strategies on the CMU-MOSEI dataset.

the most, while audio and video alone perform the worst. These trends align with previous findings (Wang, Cui, and Li 2023; Gao et al. 2024). Notably, even with only Audio and Text, FINE achieves SOTA in ACC-2 and F1, demonstrating its effectiveness despite being designed for tri-modal input.

Role of Key Components Table 6 presents the ablation results that verify the effectiveness of each key component in our proposed FINE framework.

MoQ plays an essential role in early-stage feature extraction. It dynamically selects from lightweight Q-Former experts to model fine-grained affective cues from multiple perspectives. Replacing MoQ with a standard Transformer Encoder (TE) or lightweight alternatives such as LSTM, TCN, or MLPs results in notable performance degradation. These baseline networks struggle to model token-level interactions effectively and are less robust to padding noise. Furthermore, their output structures are less compatible with the downstream FTRE module, leading to suboptimal integration and sentiment recognition.

FTRE is critical for disentangling modality-shared and modality-unique features while filtering irrelevant noise via task-aware constraints. Replacing FTRE with MISA (Hazarika, Zimmermann, and Poria 2020), which performs modality-invariant and modality-specific decomposition, leads to significant drops in ACC-2 and F1. To further investigate the role of mutual information estimation, we ablate the task-aware constraint and observe that removing task relevance causes only marginal gains over existing methods. In contrast, preserving only task-aware components without disentanglement also fails to improve performance. Interestingly, even without task relevance, our MI-based decomposition still achieves SOTA-level results, suggesting that mitigating redundancy and enhancing complementarity alone is highly effective. When we isolate only shared or only unique branches, performance degrades drastically, confirming that the disentanglement of both is essential to balanced representation learning.

DCQ contributes by modeling long-term temporal dynamics and reducing incidental noise. Compared to standard Supervised Contrastive Learning (SCL) (Khosla et al. 2020), DCQ consistently improves performance, owing to its class-wise sub-queue design and memory of historical task-relevant embeddings. Furthermore, incorporating ACCon into the contrastive loss further improves results, highlighting the advantage of modeling label-aware similarity in continuous sentiment spaces.

Configs	ACC-2	F1	ACC-7
w/o $\mathcal{L}_{UP}$	86.76	86.63	51.58
Role of " Mixture of Q-Formers "			
w/o MoQ	86.96	86.90	42.93
w/o MoQ + TE	86.21	86.01	50.76
w/o MoQ + TCN	86.65	86.71	54.97
w/o MoQ + LSTM	86.21	86.22	52.01
w/o MoQ + MLPs	84.87	85.01	50.98
Role of " Factorized Task-Relevant Encoder "			
w/o I_{str}, I_{utr}	86.79	86.64	54.95
w/o I_{sha}, I_{uni}	86.32	86.40	52.80
w/o I_{sha}, I_{str}	85.31	85.47	53.19
w/o I_{uni}, I_{utr}	85.86	85.95	52.29
w/o I_{sha}	84.40	83.76	52.05
w/o FTRE	86.16	85.82	53.12
w/o FTRE + MISA	86.98	86.96	54.32
Role of " Dynamic Contrastive Queue "			
w/o DCQ	86.41	86.31	48.42
w/o DCQ + ACCon	86.76	86.64	51.43
w/o DCQ + SCL	86.65	86.60	48.32
DCQ (SCL)	86.85	86.90	50.30
FINE	**87.70**	**87.68**	**55.59**

Table 6: Ablation studies for the key components on the CMU-MOSEI dataset.

These ablation studies comprehensively validate the effectiveness and synergy of the FINE framework's components. They emphasize the core value of feature disentanglement, and demonstrate that reducing redundancy, enhancing modality complementarity, and improving task relevance are all crucial—and interdependent—factors in achieving robust and expressive multimodal sentiment representations.

Conclusion

In this paper, we propose FINE, a multimodal sentiment analysis framework based on mutual information estimation. At its core, FINE employs a Factorized Task-Relevant Encoder to disentangle input features into shared and unique branches while suppressing task-irrelevant noise. This strategy reduces redundancy, enhances modality complementarity, and improves alignment with sentiment prediction. To support this process, we introduce a Mixture of Q-Formers for early fine-grained feature extraction and a Dynamic Contrastive Queue for long-term pattern modeling. Together, these modules enable FINE to construct robust and expressive representations under heterogeneous and asynchronous multimodal conditions. Experiments on multiple benchmarks demonstrate FINE's superior performance and confirm the importance of disentanglement and task relevance in multimodal sentiment analysis.

Acknowledgments

This work was supported by National Natural Science Foundation of China 62376255.

References

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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)*, 2236–2246. Melbourne, Australia: Association for Computational Linguistics.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. CLUB: a contrastive log-ratio upper bound of mutual information. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20029–20038.

Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; and Hussain, A. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91: 424–444.

Gao, Z.; Hu, D.; Jiang, X.; Lu, H.; Shen, H. T.; and Xu, X. 2024. Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 9650–9659. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, 6–15. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384810.

Hasan, M. K.; Rahman, W.; Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 1122–1131. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

Jiang, X.; Xu, X.; Zhang, J.; Shen, F.; Cao, Z.; and Shen, H. T. 2024. SDN: Semantic Decoupling Network for Temporal Language Grounding. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5): 6598–6612.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li, W.; Zhou, H.; Yu, J.; Song, Z.; and Yang, W. 2024. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37: 59808–59832.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6631–6640.

Liang, P. P.; Deng, Z.; Ma, M. Q.; Zou, J. Y.; Morency, L.-P.; and Salakhutdinov, R. 2023. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 32971–32998. Curran Associates, Inc.

Liang, P. P.; Lyu, Y.; Fan, X.; Tsaw, J.; Liu, Y.; Mo, S.; Yogatama, D.; Morency, L.-P.; and Salakhutdinov, R. 2022. High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning. *Transactions on Machine Learning Research*.

Lin, R.; and Hu, H. 2024. Dynamically Shifting Multimodal Representations via Hybrid-Modal Attention for Multimodal Sentiment Analysis. *IEEE Transactions on Multimedia*, 26: 2740–2755.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Bagher Zadeh, A.; and Morency, L.-P. 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)*, 2247–2256. Melbourne, Australia: Association for Computational Linguistics.

Liu, Z.; Wu, M.; Cao, W.; Chen, L.; Xu, J.; Zhang, R.; Zhou, M.; and Mao, J. 2017. A facial expression emotion recognition based human-robot interaction system. *IEEE CAA J. Autom. Sinica*, 4(4): 668–676.

Ma, F.; Zhang, Y.; and Sun, X. 2023. Multimodal Sentiment Analysis with Preferential Fusion and Distance-aware Contrastive Learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1367–1372.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01): 164–172.

Mai, S.; Zeng, Y.; and Hu, H. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25: 4121–4134.

Melville, P.; Gryc, W.; and Lawrence, R. D. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 1275–1284. New York, NY, USA: Association for Computing Machinery. ISBN 9781605584959.

Morency, L.-P.; Mihalcea, R.; and Doshi, P. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, 169–176. New York, NY, USA: Association for Computing Machinery. ISBN 9781450306416.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34: 14200–14213.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Petrovica, S.; Anohina-Naumeca, A.; and Ekenel, H. K. 2017. Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science*, 104: 437–444.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 439–448.

Pramanick, S.; Roy, A.; and Patel, V. M. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3930–3940.

Rahman, W.; Hasan, M. K.; Lee, S.; Bagher Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369. Online: Association for Computational Linguistics.

Sánchez-Rada, J. F.; and Iglesias, C. A. 2019. Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion*, 52: 344–356.

Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; and Pantic, M. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65: 3–14.

Sun, H.; Wang, H.; Liu, J.; Chen, Y.-W.; and Lin, L. 2022. CubeMLP: An MLP-based Model for Multimodal Sentiment Analysis and Depression Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 3722–3729. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. Florence, Italy: Association for Computational Linguistics.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *International Conference on Representation Learning*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Wang, Y.; Cui, Z.; and Li, Y. 2023. Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 21968–21977.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 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)*, 7617–7630. Toronto, Canada: Association for Computational Linguistics.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 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*, 3718–3727.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 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, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. Copenhagen, Denmark: Association for Computational Linguistics.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv preprint arXiv:1606.06259*.

Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J. T.; and Peng, X. 2023. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zhao, B.; Qu, X.; Kang, Z.; Peng, J.; Xiao, J.; and Wang, J. 2025. ACCon: Angle-Compensated Contrastive Regularizer for Deep Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22750–22758.