

Supplementary Materials: Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis

Anonymous Authors

1 OVERVIEW OF SUPPLEMENTARY MATERIALS

In this supplementary material, we delve deeper into the intricacies of our proposed EUAR method, providing extensive details that were not feasible to include within the confines of our primary submission due to space limitations. This supplementary mainly contains the following contents:

- Section 1: Here, we present additional experimental details, encompassing comprehensive statistics of all utilized datasets and an expanded array of hyperparameter settings employed during both training and testing phases. Within this section, we also offer a more intricate exposition of the training and testing methodologies integral to our EUAR method.
- Section 2: A comprehensive understanding can be achieved with access to the full set of experimental results and supplementary ablative analysis.
- Section 3: Additional visualization results and qualitative analysis, including a more extensive examination of classification outcomes under various noise conditions, are provided for a comprehensive understanding.
- Section 4: Limitation discussions, we analyze the shortcomings of our work and highlight potential directions for future research.

For more details, please refer to our anonymous implementation code <https://anonymous.4open.science/r/EUAR-7BF6>.

2 MORE IMPLEMENTATION DETAILS

Datasets. As described in the formal paper, we adopt 5 multimodal datasets to evaluate our method. Here we provide more details of the datasets as follows:

- *CMU-MOSI* [6]: The CMU-MOSI dataset comprises 2,199 human-annotated speech video clips, with each clip associated with an emotional intensity score ranging from -3 to 3. Its training, validation, and test sets consist of 1,281, 229, and 685 samples, respectively. Each sample includes visual, audio, and text descriptions. This dataset is collected from social media and has been annotated by people who invested their time and effort. Due to its collection from social media, it inevitably contains some unavoidable noise.
- *CMU-MOSEI* [1]: The CMU-MOSEI dataset is an extensive collection comprising 22,856 movie review segments sourced from YouTube. It's a substantial dataset tailored for Multimodal Sentiment Analysis (MSA) tasks, featuring training, validation, and test sets consisting of 16,265, 1,869, and 4,643 samples, respectively. This dataset includes labels for both emotion and sentiment. Despite being a multi-label dataset, for this particular task, we exclusively focus on the sentiment labels.

- *MVSA-Single* [4]: The MVSA-Single dataset is a bimodal dataset created for Multimodal Sentiment Analysis (MSA) tasks. It consists of pairs of images and text gathered from social media platforms. Both the image and text components in this dataset contain different levels of noise. Sentiment is categorized into three classes: positive, neutral, and negative. Classification is performed using both the images and their associated textual descriptions.
- *NYU Depth v2* [5]: We utilize RGB images and depth images sourced from the NYU Depth v2 dataset as input for predicting scene categories. Following the standard protocol [7], we divide the 1,449 samples into a training set and a test set with a ratio of 795 to 654. We preserve the nine most common scenarios and relabel the remaining categories as "others." In this dataset, there are two modalities: RGB images and depth images. We explored this experiment to delve into more modalities. Additionally, this dataset contains some data noise, which conveniently facilitates our experiment.

Moreover, following [7], we also use noisy datasets to evaluate the robustness of multimodal fusion in our method, Noisy NYU Depth v2, where Gaussian noise and Salt-Pepper noise are jointly considered. For clarity, here we summarize the overall statistics of the datasets above in Table 1.

Metrics. For CMU-MOSI and CMU-MOSEI, in line with previous research [3], we employed the F1 score for binary classification, seven-class accuracy, and Pearson correlation coefficient. For binary classification, we grouped the predicted sentiment intensity scores into positive and negative classes and reported the F1 score. For the seven-class classification, we rounded the predicted continuous sentiment intensity values to -3, -2, -1, 0, 1, 2, 3, and reported the accuracy.

For the MVSA-Single and NYU Depth v2 classification datasets, we reported their accuracy and F1 score.

To test the robustness of the EUAR method, following prior research [7], we introduced Gaussian noise or Salt-Pepper noise to the test data of NYU Depth v2. For Gaussian noise, we set the mean to 0 and the variance to ϵ for the visual modality. To ensure a fair comparison, we conducted our method on this dataset ten times using different random number seeds. We averaged the experimental results and reported both the mean and standard deviation.

For a fair comparison, we conducted 10 experiments using different random number seeds and reported the average accuracy.

Experimental Settings. To ensure a fair comparison, we align the backbone of our feature extraction with existing state-of-the-art methods. For the trimodal dataset, we employ FACET, COVAREP, and BERT as feature extractors for the visual, audio, and text modalities, respectively. The extracted feature dimensions for visual, audio, and text are 64, 64, and 768, respectively. For the MVSA-Single dataset, we utilize ResNet-152 and BERT as feature extractors for visual and text features, respectively. Regarding the NYU Depth v2

Table 1: The statistics of the adopted multimodal datasets. Note that the CMU-MOSI and CMU-MOSEI datasets, which are used for regression-based tasks, are labeled with sentiment strength from -3 to 3.

Datasets	Noise	Modality	Train/Val/Test	Class	Task
CMU-MOSI	-	Video, Text, Audio	1281/229/685	7	Regression
CMU-MOSEI	-	Video, Text, Audio	16265/1869/4643	7	Regression
MVSA-Single	-	RGB, Text	1555/518/519	3	Classification
NYU Depth v2	-	Depth, RGB	795/414/654	10	Classification
NYU Depth v2	Gaussian, Salt-Pepper	Depth, RGB	795/414/654	10	Classification

dataset, which utilizes both depth and RGB images as two modalities, we employ ResNet152 as the feature extractor for both modalities. We set x , y , and z to be $1e-3$, $1e-5$, and $1e-3$, respectively. This enabled us to achieve the results reported in the paper. Additionally, we conducted sensitivity analysis on the hyperparameters, the results of which are shown in Fig. 5. We conducted a hyperparameter sensitivity analysis experiment on the CMU-MOSI dataset. We fixed λ and β respectively and varied the values of the other two parameters, testing the model's performance and reporting the Acc2 metric. In the left figure, we fixed λ , and when changing the value of β , the model's performance fluctuated significantly, indicating that our model is highly sensitive to the hyperparameter β , and special attention should be paid to the setting of β during training. Meanwhile, in the right figure, we fixed the value of β and changed the other two hyperparameters, and the fluctuation in model performance was not significant. From these results, it can be noted that during training, special attention should be paid to the trade-off weight of auxiliary loss.

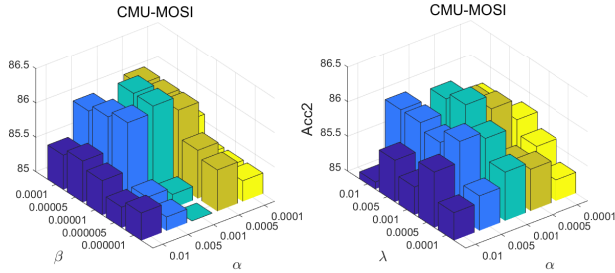


Figure 1: Hyperparameter sensitivity analysis. We conducted tests on different hyperparameter settings to observe the model's Acc2 metric on the CMU-MOSI dataset.

Training and Testing Details We outlined our training procedure in Algorithm 1. Meanwhile, in Algorithm 2, we elucidated our testing procedure to provide a better understanding of our algorithm.

3 ADDITIONAL FURTHER ANALYSIS

Additional Experimental Results with standard deviation Due to space constraints in the main paper, we did not report the standard deviation of the test results on the noisy dataset. Therefore, in Table 2, we supplemented the standard deviation metric. We conducted additional experiments on the noisy dataset NYU Depth V2 using different random seeds and reported the mean and standard deviation of the model results. From the results, it can be observed

Algorithm 1: Training procedure of the EUAR

Data: Video clips or text-image pairs
Result: Updated model parameters

- 1 Extract features from each unimodal data to obtain X_i^m
- 2 **while** Training **do**
- 3 For each unimodal input, obtain X_i^m through encoder backbone;
- 4 Feed the input into the Gate in the MoE framework, producing routing weights G_j^m ;
- 5 Select the expert to be used based on the routing weights and use two separate fully connected layers to predict μ_i^m and σ_i^m , respectively;
- 6 Calculate the unimodal representation h_i^m :

$$h_i^m = \sum_j G_j^m(x_i^m) E_j^m(x_i^m);$$
- 7 Calculate the auxiliary loss: $L_{aux}^m = \frac{1}{N} \sum_j R_j^m P_j^m$;
- 8 Calculate the U-loss: $L_u^m = \frac{1}{N} \sum_j \sigma_j^{m2} G_j(x^m)$;
- 9 Calculate the KL-loss:

$$L_{kl}^m = -\frac{1}{2} (1 + \log \sigma_i^{m2} - \mu_i^{m2} - \sigma_i^{m2});$$
- 10 Concatenate the unimodal representations to obtain \hat{h}_i ;
- 11 Pass \hat{h}_i through the final fully connected layer to obtain regression or classification results \hat{y}_i ;
- 12 Calculate L_{task} ;
- 13 Calculate L_{total} ;
- 14 Update model parameters with Adam optimizer;

that our model not only achieved superior performance in terms of average accuracy but also outperformed other state-of-the-art models in terms of standard deviation. It is noteworthy that as the noise intensity increases, the superiority of our model in terms of standard deviation becomes more pronounced, demonstrating the robustness of our approach. This confirms the superiority of our method in handling noisy data.

Additional Ablation Experiment on Loss To demonstrate the effectiveness of the uncertainty loss we proposed, we conducted additional experiments on the noisy dataset NYU Depth V2. We tested the model without U-loss on the noisy dataset and compared it with our full model. The experimental results are shown in Figure X. From the figure, it can be seen that our complete EUAR method performs significantly better under various noise conditions. It is noteworthy that the greater the intensity of the noise, the greater the advantage of our complete method. This fully demonstrates the effectiveness of the proposed U-loss in handling noisy data.

Specifically, when the intensity of Gaussian noise and salt-and-pepper noise is 10, our complete method outperforms by nearly 2% in accuracy, highlighting the superiority of our approach.

Algorithm 2: Testing procedure of the EUAR

Data: Video clips or text-image pairs

Result: Regression or classification results

- 1 For each unimodal input, obtain X_i^m through encoder backbone;
 - 2 Feed the input into the Gate in the MoE framework, producing routing weights G_j^m ;
 - 3 Select the expert to be used based on the routing weights and use two separate fully connected layers to predict μ_i^m and σ_i^m , respectively;
 - 4 Calculate the unimodal representation h_i^m :

$$h_i^m = \sum_j^N G_j^m(x_i^m) E_j^m(x_i^m);$$
 - 5 Concatenate the unimodal representations to obtain \hat{h}_i ;
 - 6 Pass \hat{h}_i through the final fully connected layer to obtain regression or classification results \hat{y}_i ;
-

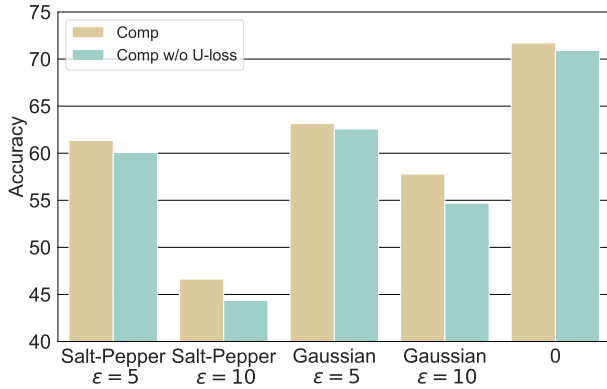


Figure 2: More t-SNE results.We conducted ablation experiments by removing U-loss under various types and degrees of noise conditions and compared it with our complete method.

Additional Ablation Experiment on Expert Number Our approach introduces the Mixture of Experts framework for the first time in the task of multimodal sentiment analysis. To demonstrate the superiority of this dynamic network, we conducted tests on modalities missing when the number of experts in MoE is decreased, comparing it with our complete method. We conducted experiments on both CMU-MOSI and CMU-MOSEI datasets to observe the impact of varying the number of experts on model performance. The results are shown in Figure 3. From the figure, it can be observed that our method consistently outperforms in terms of the Acc2 metric under various expert number scenarios. This clearly demonstrates the correctness of introducing the MoE framework into multimodal sentiment analysis. Our method achieves good performance across different numbers of experts. However, as the number of experts increases, our model’s performance also improves, demonstrating

the effectiveness of the proposed approach. It is noteworthy that our method demonstrates a significant performance advantage as the number of experts increases from 2 to 32. These results demonstrate that our method not only achieves good performance in noisy data scenarios but also performs well across different numbers of experts. Furthermore, this performance advantage increases with the increase in the number of experts.

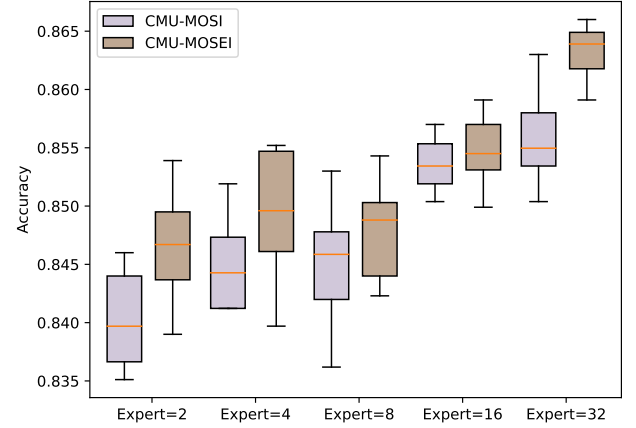


Figure 3: Expert Number Ablation results.We tested the model performance using different numbers of experts on both the CMU-MOSI and CMU-MOSEI datasets.

4 ADDITIONAL VISUALIZATION

In order to visually demonstrate the effectiveness of our method, we conducted additional visualizations on EUAR. We conducted feature visualization under more noisy conditions, showcasing additional t-SNE visualization results. Simultaneously, we conducted more qualitative analysis, comparing our method with state-of-the-art approaches.

Visualization of Joint Representations Similar to the same section of our paper, to fully explicate the superiority of proposed EUAR, we conduct more experiment of t-SNE visualization, which project joint representations into 2D space. Followed our method in the paper, we employ the t-SNE to visualize the learned joint representations of simple concatenation, MLP, and complete EUAR for a quantitative comparison. Moreover, in this supplementary material, we visualize the rest of the missing modalities combinations: text and audio, only text, only audio, only vision, audio and vision. As shown in Figure, comparing to simple concatenation, MLP model can distinguish joint representations of different sentimental labels to some degree, but it is still hardly tell the difference between the three clusters, especially in the combinations of only audio, only video, video and audio. It is worth noting that some of the representations are confused by the MLP model, clustering in the wrong labels. On the other hand, the complete EUAR has more compact and distinguishing clusters. However, because of the the inherent characteristic of CMU-MOSI, that the textual modality mainly contribute to provide more valid information than the other two modalities, when the text modality is missing, the joint representations of audio and vision contain more noise, making it

Table 2: Noisy NYU v2 Depth. We conducted 10 experiments using different random seeds, reporting the average and standard deviation of model accuracy.

Method	Clean	Salt-Pepper Noise		Gaussian Noise	
	Acc@ $\epsilon = 0$	Acc@ $\epsilon = 5$	Acc@ $\epsilon = 10$	Acc@ $\epsilon = 5$	Acc@ $\epsilon = 10$
Concat	70.44 \pm 0.89	57.98 \pm 2.08	44.51 \pm 2.91	59.97 \pm 2.14	53.20 \pm 3.50
Late fusion	69.16 \pm 0.68	56.27 \pm 2.40	41.22 \pm 2.78	59.63 \pm 2.44	51.99 \pm 3.11
Align	70.31 \pm 1.28	57.54 \pm 2.50	43.01 \pm 2.66	59.47 \pm 1.84	51.74 \pm 3.41
MMTM	71.04 \pm 0.41	59.45 \pm 1.38	44.59 \pm 2.49	60.37 \pm 2.61	52.28 \pm 3.77
TMC [2]	71.06 \pm 0.76	59.34 \pm 1.03	44.65 \pm 2.30	61.04 \pm 1.66	53.36 \pm 2.76
QMF [7]	70.09 \pm 0.97	58.50 \pm 2.05	45.69 \pm 2.79	61.62 \pm 1.84	55.60 \pm 2.09
RSMF (Ours)	71.71 \pm 0.87	61.35 \pm 1.32	61.35 \pm 2.01	63.15 \pm 1.67	57.79 \pm 1.97

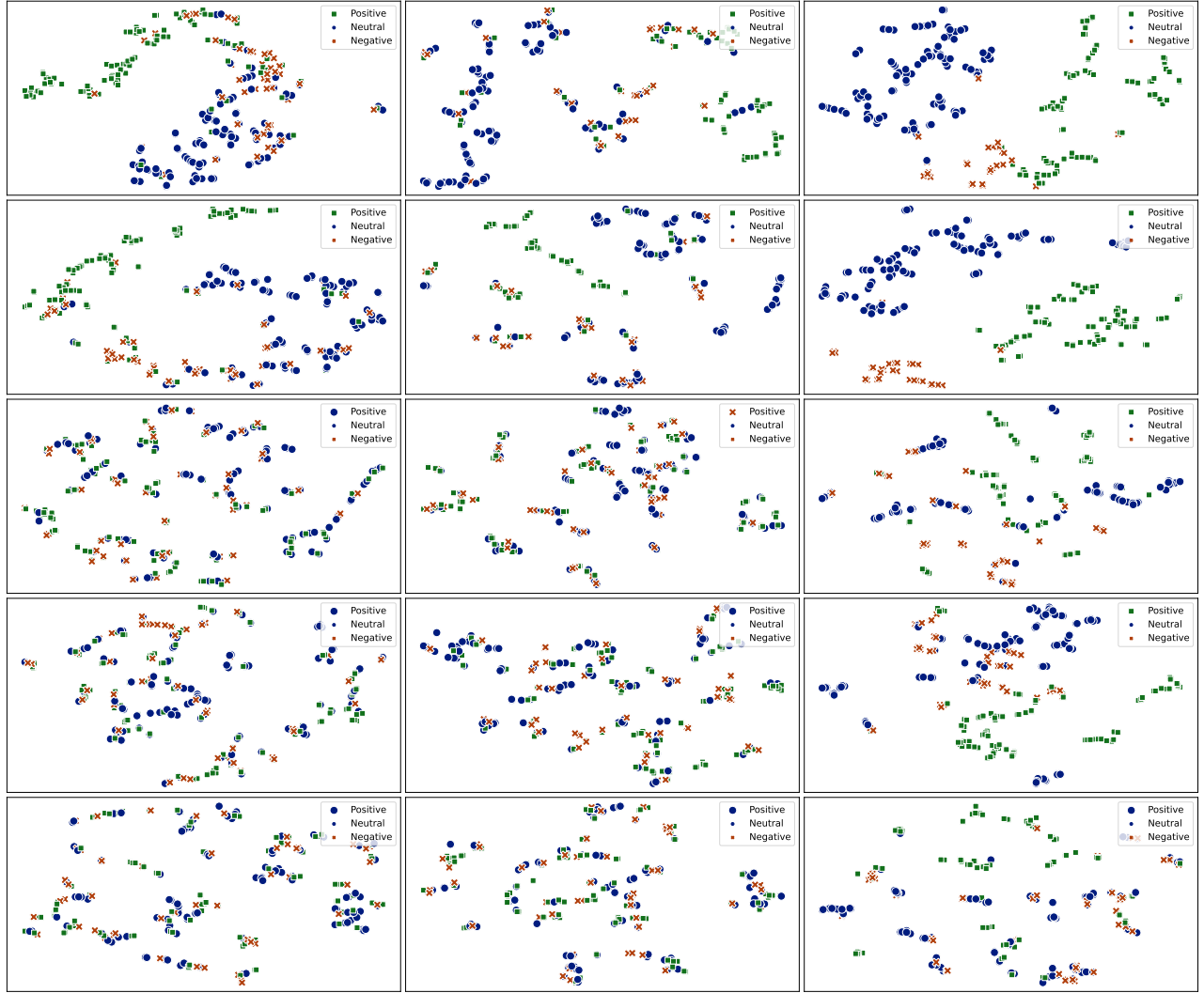


Figure 4: More t-SNE results. We supplemented additional feature visualizations for cases of missing modality, comparing our complete method with the ablated one.

Modality (L, V, A)	Text Modality	Video Modality	Audio Modality	Predictions	
				EAUR (Ours)	DiCMoR (ICCV'23)
	Missing		Missing	positive ✓	neutral ✗
	Missing		Missing	negative ✓	neutral ✗
	Missing		Missing	negative ✓	positive ✗
	Missing	Missing		negative ✓	positive ✗
	Missing	Missing		positive ✓	negative ✗
	Missing	Missing		negative ✓	neutral ✗
	Missing			negative ✓	positive ✗
	Missing			positive ✓	neutral ✗

Figure 5: More Qualitative Analysis results. We gathered additional qualitative analysis results and visualized them for comparison with the current state-of-the-art methods.

hard to classify. This is the reason we suspect why the results of clusters in the last three rows of the figure are not so distinctive as the rest of the missing combinations. Nevertheless, our proposed model manages to learn the joint representations from those two modalities by excluding noise within these modalities, and can still outperform the other ablated models. Consequently, the t-SNE visualization indicate that with the help of both MoE structure and

proposed U-loss, EUAR is more capable of learning a representative joint representation.

Qualitative Analysis Due to the limited space of our paper, we exhibit more qualitative analysis on missing modalities scenarios on CMU-MOSI dataset. As shown in figure above, except for the combinations whose textual modalities are available, we also conduct following combinations whose texts are missing, namely

only video, only audio and both video and audio. These three combinations, due to the unavailability of textual modality, contain more noise and invalid information comparing to combinations we present in the paper, resulting in more challenging classification. Followed our method in the paper, we select several typical cases from testing set of CMU-MOSI. Also, we use rectangles with red dash lines to represent missing modalities. It is obvious to observe that our proposed EUAR method consistently produce correct classification, while DiCMoR is still unable to conduct the MSA task accurately. Ulteriorly, quantitative visualization of above combinations show that the transcendent robustness and the ability of noise resistance of our proposed EUAR.

5 LIMITATION DISCUSSION

While our method has achieved state-of-the-art results, there are still some limitations worth noting. Firstly, our method primarily relies on aleatoric uncertainty of the data, potentially overlooking another form of uncertainty known as epistemic uncertainty. Existing methods for quantifying epistemic uncertainty are often computationally expensive and involve a large number of parameters. Therefore, it is imperative for future research to explore more cost-effective and lightweight methods for quantifying epistemic uncertainty and integrating it into feature enhancement processes.

Secondly, we have observed that on datasets with high data quality, such as CMU-MOSEI, our method, while achieving state-of-the-art performance, shows limited improvement. When dealing

with high-quality data, our method struggles to effectively handle the data. Hence, there is a need in the future to develop a more comprehensive data processing method that can demonstrate superior performance regardless of the dataset quality. At the same time, we placed greater emphasis on individual data handling by different experts without making significant efforts to coordinate between them. In the future, we will explore more methods to facilitate collaboration among different experts to fully leverage the power of the entire framework.

REFERENCES

- [1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of Meeting of the Association for Computational Linguistics*. 2236–2246.
- [2] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2020. Trusted multi-view classification. In *International Conference on Learning Representations*.
- [3] Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia* (2022).
- [4] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on MultiMedia Modeling*. 15–27.
- [5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*. 746–760.
- [6] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* (2016), 82–88.
- [7] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. 2023. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*. 41753–41769.