Github Supplemental Material for "Two-Stage Dynamic Fusion Framework for Multimodal Classification Tasks"

Shoumeng Ge
School of Management, Harbin Institute of Technology, 23B910001@stu.hit.edu.cn
Ying Chen*
School of Management, Harbin Institute of Technology, yingchen@hit.edu.cn (*corresponding author)

Appendix F. Additional Discussion

F.1. Extending to three modalities

Our method is a late fusion method, making it easily extensible to incorporate more modalities. Here, we apply it to two multimodal sentiment analysis datasets with three modalities i.e., CMU-MOSI (Pérez-Rosas et al. 2013) and CMU-MOSEI (Wöllmer et al. 2013). Each dataset contains visual, audio, and text modalities.

CMU-MOSI dataset was developed by Pérez-Rosas et al. (2013). It consists of 93 vlogs collected from the YouTube website. These types of videos typically feature a single speaker, aged between 20 and 30 years, with 41 videos expressed by females and the rest by males, all conveying opinions in English. One advantage of this dataset is its ability to handle diverse content, including noise. Additionally, it is important to note that all videos were recorded in different settings, resulting in variations. Some users used high-tech microphones and cameras, while others used less professional recording equipment. Moreover, the distance between users and the camera, as well as differences in background and lighting conditions, contribute to these variations. The videos maintain their original resolution without any quality enhancements. The labels in this dataset are divided into: strongly positive (+3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), and strongly negative (-3).

CMU-MOSEI dataset was developed by Wöllmer et al. (2013). It is a larger-scale dataset consisting of 3,228 videos, comprising 22,777 utterances from over 1,000 YouTube

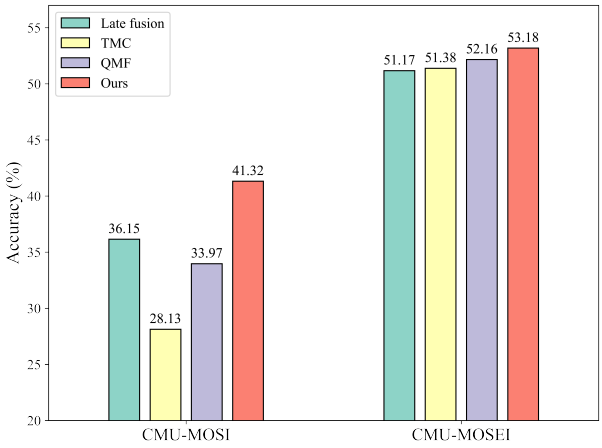


Figure F.1 Comparison of classification accuracy for CMU-MOSI and CMU-MOSEI datasets.

users (57% male and 43% female). These videos cover 250 different topics, with the three most common topics being comments (16.2%), debates (2.9%), and consultations (1.8%). Like the CMU-MOSI dataset, CMU-MOSEI can handle diversity and includes noise. Each utterance in the dataset is labeled as one of the following seven sentiment score: strongly positive (+3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), and strongly negative (-3).

We employed a transformer encoder as the backbone network and compared three late fusion methods: Late Fusion, TMC, and QMF, as shown in Figure F.1. The results on both datasets indicate that our method achieves the best performance even with three modalities and exhibits significant superiority on the CMU-MOSI dataset. This demonstrates the effectiveness and scalability of the proposed approach.

F.2. Comparison with existing two-stage multimodal learning methods

To better demonstrate the novelty and performance of the proposed two-stage method, we compare it with four latest two-stage multimodal methods, including:

- IMF (Li et al. 2023b): A multimodal link prediction model based on knowledge graphs, which employs bilinear fusion to integrate multimodal information in the first stage. In the second stage, it combines the results of different models using fixed weights;
- TSIEN (Tan et al. 2024): A multi-view multi-label classification model. In the first stage, mutual information is used to train view-specific classifiers, extracting task-relevant information. In the second stage, an autoencoder-based mutual information extraction framework is used to perform weighted fusion of the classifier inputs;

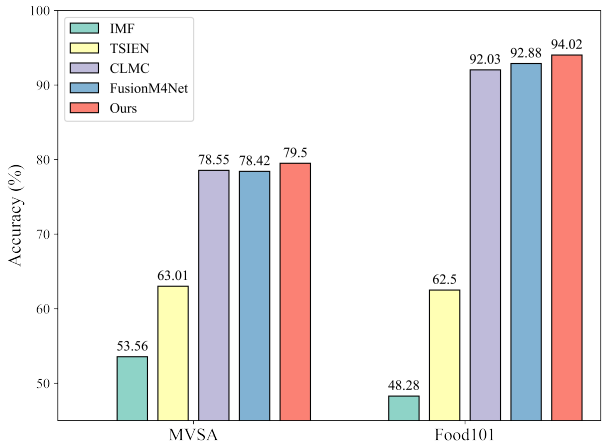


Figure F.2 Accuracy of two-stage methods on MVSA and Food101 datasets.

- CLMC (Mandal et al. 2024): A two-step training process for multimodal classification of crisis related tweets. The first step leverages contrastive learning to extract features, while the second stage involves task-specific fine-tuning;
- FusionM4Net (Tang et al. 2022): A two-stage multimodal multi-label skin lesion classification method. In the first stage, decision-level fusion integrates image information, while in the second stage, support vector machine clusters fuse non-image and image modalities.

We adapt these methods to multimodal image-text classification tasks, using ResNet152 and BERT as modality encoders. The comparison is conducted on the MVSA and Food101 datasets, with the results shown in Figure F.2. The results reveal that IMF and TSIEN achieve relatively low accuracy, primarily because their designs are tailored for link prediction and multi-view multi-label classification tasks, respectively. CLMC and FusionM4Net perform much better than IMF and TSIEN but still fall short of our proposed method. Thus, our method provides advantages over existing two-stage multimodal learning methods.

F.3. Impact of different regression methods at the second stage

We compare the performance of different regression models in the second stage of multimodal fusion, including linear regression (LR), random forests (RF), support vector regression (SVR), and evidence-based multilayer perceptrons (EMLP), to adaptively generate fusion weights based on the learned uncertainty from the first stage, as shown in Table F.1. We repeat each experiment ten times. Overall, EMLP demonstrates the better performance than the existing ones in 9 out of 12 scenarios, establishing itself as the

Table F.1 Results of four datasets using different regression models in the second stage.

Method	MVSA			CrisisMMD		
	$\sigma=0,p=0$	$\sigma=5,p=0.5$	$\sigma=10,p=1$	$\sigma=0,p=0$	$\sigma=5,p=0.5$	$\sigma=10,p=1$
LR	79.46±1.54	73.99±1.58	67.13±1.16	87.67±0.46	79.77±1.04	73.73±0.65
RF	79.21±1.25	73.76±1.54	**67.65±1.17**	87.55±0.42	79.96±0.90	**74.89±0.53**
SVR	**79.63±1.47**	73.66±1.76	67.57±1.20	87.50±0.50	79.53±0.87	72.94±0.44
EMLP	79.50±1.51	**74.32±1.78**	67.50±1.49	**87.71±0.40**	**80.26±0.91**	74.83±0.48
	N24News			Food101		
	$\sigma=0,p=0$	$\sigma=5,p=0.5$	$\sigma=10,p=1$	$\sigma=0,p=0$	$\sigma=5,p=0.5$	$\sigma=10,p=1$
LR	79.89±0.21	66.76±0.69	55.28±1.57	93.86±0.11	76.87±0.33	64.24±0.21
RF	79.79±0.24	67.42±0.69	58.06±0.33	93.86±0.08	77.03±0.29	64.26±0.22
SVR	79.74±0.31	66.96±0.69	57.16±0.35	93.83±0.08	76.65±0.26	64.11±0.22
EMLP	**79.90±0.23**	**68.61±0.44**	**58.17±0.38**	**94.02±0.10**	**77.50±0.27**	**64.27±0.21**

top-performing method. Such results highlight the strength of neural networks in learning adaptive fusion weights, particularly in more complex and challenging datasets (N24News and Food101). LR also shows competitive results compared to the results of the first stage shown in Tables 2 and 3. As noted, LR does not require training, which is efficient for computation-demanding classification tasks. In summary, EMLP generally emerges as the most effective model for complex multimodal fusion tasks, while LR serves as a simpler yet reliable option for less demanding scenarios where computational resources are a concern.

F.4. The visualization of Remark 1

We further investigate how these uncertainty estimation methods perform under different noises. We use the results of Food101 from Table 3 for illustration, which is as shown in Figure F.3. Note that the upper one in Figure F.3 corresponds to the results presented in Table 3. The upper one is from the scenario of "$\sigma=0,p=0$", the middle one is from the scenario of "$\sigma=5,p=0.5$", and the lower one is from the scenario of "$\sigma=10,p=0.5$". For simplicity, we only use "$\sigma=0$, $\sigma=5$, and $\sigma=10$" to represent these three scenarios in Figure F.3. As observed, a strong-negative correlation is observed between the Pearson coefficient and accuracy in these three scenarios. With the level of noise increases, the correlation also increases. This may be due to the uncertainty estimation's ability to identify out-of-distribution samples more easily, thereby enhancing correlation. As the correlation increases, the performance of the multimodal model approaches the upper accuracy limit, validating the **Remark 1**.

Regarding this, there are two key factors that affect the performance of multimodal fusion models. On the one hand, the correlation between the fusion weights and $h(\cdot)$ plays a critical role. Higher correlation leads to better performance. This necessitates identifying

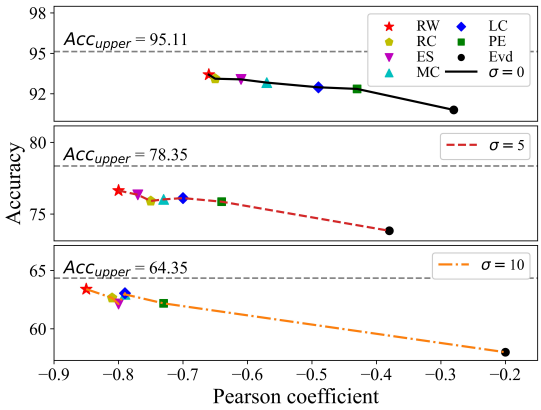


Figure F.3 **The variations between the classification accuracy and correlation of different uncertainty estimation methods with $h(\cdot)$ under the influence of noise on the Food101 dataset. The abbreviations used for the methods represent: Evidence (Evd), Predict Entropy (PE), Least Confidence (LC), Margin Confidence (MC), Energy Score (ES), Evidence-based Ratio of Confidence (RC), Regress Weight (RW).**

fusion weights that exhibit a stronger correlation with $h(\cdot)$, which can be obtained through uncertainty estimation or learning methods. On the other hand, the upper performance limit of multimodal fusion, Acc_{upper}, represents the potential of the model. When the performance of unimodal models remains constant, a higher level of diversity between modalities results in a larger Acc_{upper}, leading to a better performance of multimodal models. In contrast, with lower diversity between modalities, the lower limit of multimodal model performance is more assured.

F.5. The effects of noise type

In this section, we follow Zhang et al. (2023) and replace the Gaussian noise with salt-and-pepper noise in the images. We keep the other settings unchanged and compare the model performance on two larger datasets, N24News and Food101. We present the results in Table F.2. The results suggest that, despite changing the noise type, our model continues to outperform the counterpart models. In particular, on the N24News dataset, our model exhibits a significant advantage. The results in Table F.2 are consistent with those in Table 3, demonstrating that our model maintains its high effectiveness in terms of model accuracy and robustness under different types of noise interference compared to the counterpart models.

Table F.2 Classification accuracy comparison when 50% of the modalities are corrupted with image noise (Salt-pepper noise with a density of σ) and text noise (mask words with a probability of p) on N24News and Food101 datasets.

Method	N24News			Food101		
	$\sigma=0, p=0$	$\sigma=5, p=0.5$	$\sigma=10, p=1$	$\sigma=0, p=0$	$\sigma=5, p=0.5$	$\sigma=10, p=1$
ResNet-152	45.10±0.75	35.46±0.51	25.06±0.38	66.75±0.40	50.42±0.47	35.68±0.92
BERT	77.89±0.16	62.12±1.03	41.78±0.19	86.13±0.15	67.43±0.24	43.48±0.26
Late fusion	79.29±0.23	62.59±3.74	39.71±4.91	90.06±0.23	76.36±0.77	56.49±0.61
MMBT	79.64±0.20	63.07±1.12	42.30±0.55	91.33±0.15	76.06±0.35	54.86±0.44
TMC	76.97±0.24	60.73±5.76	42.37±7.24	89.99±0.17	78.25±0.24	60.73±0.71
ETMC	76.27±0.24	53.77±3.70	32.76±4.80	89.88±0.35	78.52±0.63	60.64±1.03
QMF	79.76±0.38	59.81±4.14	36.89±5.61	92.92±0.05	81.39±0.23	61.23±1.19
Ours	**79.90±0.23**	**65.37±0.96**	**48.77±1.31**	**93.68±0.07**	**81.83±0.07**	**62.60±0.70**

F.6. Impact of backbone network

Different backbone networks can lead to different fusion results. In this section, we change the ResNet152 backbone network to Vision Transformer (ViT) (Dosovitskiy et al. 2020) to extract the feature from the image modality since ViT has been used in multimodal learning and has achieved outstanding performance (see Zou et al. 2023). Keeping the other settings unchanged, we then train the two-stage dynamic fusion framework. Moreover, we compare our framework with ViT-based SOTA methods in the Food101 dataset, including CMA-CLIP (Liu et al. 2021), ME (Liang et al. 2022), UniS-MMC (Zou et al. 2023), and two multimodal large language models (MLLMs) (i.e., BLIP (Li et al. 2022) and BLIP-2 (Li et al. 2023a)). For MLLMs, we conduct two experiments, i.e., zero-shot and linear probe. For the linear probe, we add an additional linear layer for classification based on the settings in CLIP (Radford et al. 2021). We repeat each method five times except CMA-CLIP and ME since these two methods have already reported their results on the Food101 dataset in their papers and we just copy them. We present the results in Table F.3. As seen, our models (eRUW and RW) both outperform the existing models with this new backbone network. Interestingly, these two MLLMs have much worse performance than the other baselines. Furthermore, our method still has room for improvement as it falls short of the upper limit (Acc_{upper}). This indicates the potential for further enhancements.

F.7. Uncertainty in the decision-making process

We visualize the uncertainty distribution using a Gaussian kernel density estimation (Scott 2015). Figure F.4 illustrates the variations in the uncertainty distribution with the introduction of noise. In Figure F.4(a), the uncertainty distribution of the text unimodal model is shown before and after adding noise. Such a similar pattern is also observed in Figure

Table F.3 Performance comparison of different models after replacing the ResNet152 backbone with ViT in our framework (RWR stands for Regression-based Weight with Random Forest).

Method	Fusion	Backbone		Accuracy
		Image	Text	
CMA-CLIP	Early	ViT	Transformer	93.1
ME	Early	DenseNet	BERT	94.6
UniS-MMC	Early	ViT	BERT	$94.7_{\pm 0.1}$
BLIP Zero-Shot	Middle	ViT	BERT	$62.33_{\pm 0.11}$
BLIP Linear Probe	Middle	ViT	BERT	$84.68_{\pm 0.13}$
BLIP-2 Zero-Shot	Middle	Q-Former	LLM	$63.39_{\pm 0.1}$
BLIP-2 Linear Probe	Middle	Q-Former	LLM	$87.61_{\pm 0.12}$
eRUW(ours)	Late	ViT	BERT	$95.01_{\pm 0.13}$
R W(ours)	Late	ViT	BERT	$\mathbf{95.45}_{\pm 0.12}$
Acc_{upper}	Late	ViT	BERT	$96.40_{\pm 0.05}$

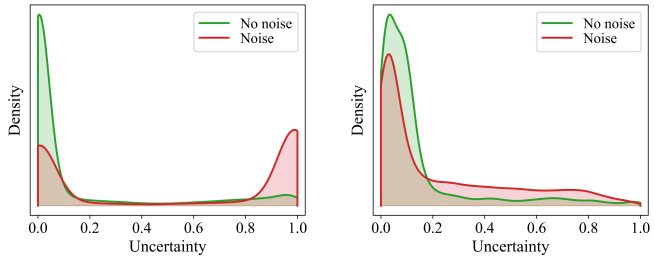
F.4(c) for the image unimodal model. These patterns clearly indicate that adding noise leads to a significant increase in uncertainty, resulting in a peak on the right side of the distribution. When adding noise to multimodal models (see Figure F.4(b)), the increase in uncertainty caused by text noise is less obvious, with no peak on the right tail. Similarly, we also do not observe a peak on the right side of Figure F.4(d) when only adding noise to the image modality. These results from Figures F.4(b) and F.4(d) demonstrate that the multimodal models built by our framework can effectively resist the influence of noise and focus more on the modality without noise.

In summary, the uncertainty dynamics in Figure F.4 elucidates why multimodal models have the capacity to resist the influence of noise and how they automatically adapt when one modality exhibits poor classification quality.

F.8. Visualization of classification

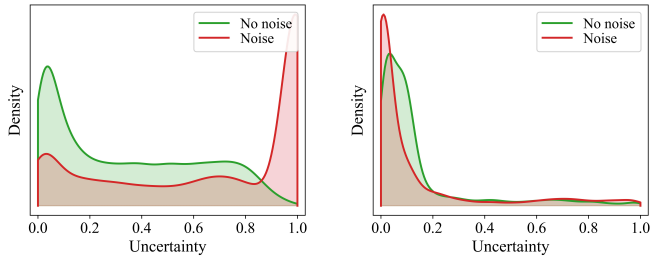
In this section, we visualize the classification results with confusion matrix for the N24News and Food101 datasets. As seen in Figure F.5, we can observe that different classes exhibit varying degrees of classification difficulty, which could be attributed to differences in the number of training samples and similarities between different classes. In Figure F.5(a), there are several classes with a high proportion of samples (a.k.a., imbalanced), making it more likely for the model to misclassify samples into these classes. This phenomenon is alleviated in Figure F.5(b), where there are fewer misclassified samples that are not visually apparent, indicating a better classification performance.

Furthermore, we follow Han et al. (2023) to explore the benefits of subjective uncertainty in decision-making. When the uncertainty u exceeds a threshold u_{max}, the prediction result for one sample is considered uncertain. We set u_{max} as 0.6 in this study. After removing the



(a) Adding noise to texts of unimodal model.

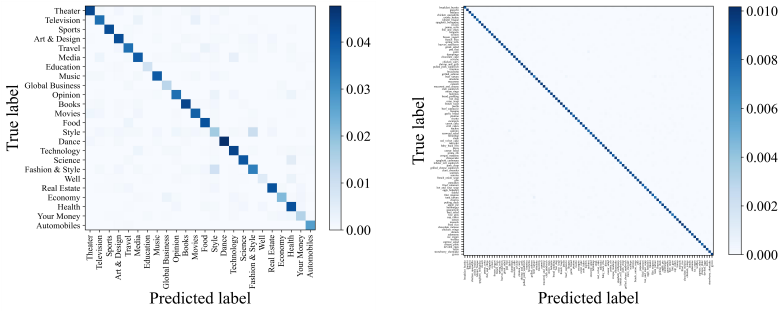
(b) Adding noise to texts of multimodal model.



(c) Adding noise to images of unimodal model.

(d) Adding noise to images of multimodal model.

Figure F.4 **The uncertainty distribution of unimodal and multimodal models after adding noise to different modalities on the Food101 dataset. Figures F.4(a) and F.4(c) demonstrate the strong uncertainty perception capability of unimodal models, while Figures F.4(b) and F.4(d) illustrate our multimodal models' ability to resist noise. They reduce the impact of low-quality modalities to ensure the reliability of predictions.**

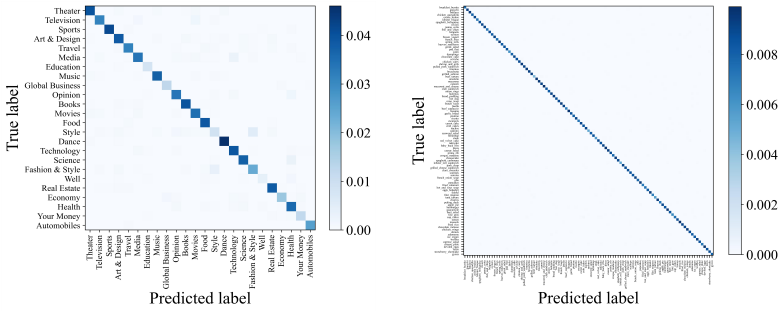


(a) N24News.

(b) Food101.

Figure F.5 **Confusion matrices of N24News and Food101 datasets.**

uncertain samples, we can obtain a subjective confusion matrix as shown in Figure F.6. For the N24News case, the classification results show significant improvements, comparing Figure F.5(a) with F.6(a). This indicates that the confusion matrix becomes more consolidated after the uncertain samples are discarded. Additionally, observing the classification accuracy of each class in Figures F.5(a) and F.6(a), we find out that changes in the confu-



(a) N24News. (b) Food101.

Figure F.6 Subjective confusion matrices of N24News and Food101 datasets.

sion matrix are less noticeable for some classes. This further denotes the variations in the classification difficulty levels across different classes. On the other hand, for the Food101 dataset, the changes in the confusion matrix are less prominent, comparing Figure F.5(b) to F.6(b). This is primarily due to the strong performance of the multimodal model on this dataset.

Moreover, the results from Figures F.5 and F.6 reveal that the use of subjective uncertainty can significantly reduce the misclassification rate, decreasing from 0.204 to 0.124 in the N24News dataset and from 0.062 to 0.016 in the Food101 dataset. This demonstrates that our model can effectively conduct the classification tasks that are challenging to recognize through subjective uncertainty.

References

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* .

Han Z, Zhang C, Fu H, Zhou JT (2023) Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence* 45(2):2551–2566.

Li J, Li D, Savarese S, Hoi S (2023a) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, 19730–19742 (PMLR).

Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International conference on machine learning*, 12888–12900 (PMLR).

Li X, Zhao X, Xu J, Zhang Y, Xing C (2023b) Imf: interactive multimodal fusion model for link prediction. *Proceedings of the ACM Web Conference 2023*, 2572–2580.

Liang T, Lin G, Wan M, Li T, Ma G, Lv F (2022) Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15492–15501.

Liu H, Xu S, Fu J, Liu Y, Xie N, Wang CC, Wang B, Sun Y (2021) Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562* .

Mandal B, Khanal S, Caragea D (2024) Contrastive learning for multimodal classification of crisis related tweets. *Proceedings of the ACM on Web Conference 2024*, 4555–4564.

Pérez-Rosas V, Mihalcea R, Morency LP (2013) Utterance-level multimodal sentiment analysis. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 973–982.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763 (PMLR).

Scott DW (2015) *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons).

Tan X, Zhao C, Liu C, Wen J, Tang Z (2024) A two-stage information extraction network for incomplete multi-view multi-label classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15249–15257.

Tang P, Yan X, Nan Y, Xiang S, Krammer S, Lasser T (2022) Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. *Medical Image Analysis* 76:102307.

Wöllmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, Morency LP (2013) Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28(3):46–53.

Zhang Q, Wu H, Zhang C, Hu Q, Fu H, Zhou JT, Peng X (2023) Provable dynamic fusion for low-quality multimodal data. *arXiv preprint arXiv:2306.02050* .

Zou H, Shen M, Chen C, Hu Y, Rajan D, Chng ES (2023) Unis-mmc: Multimodal classification via unimodality-supervised multimodal contrastive learning. *arXiv preprint arXiv:2305.09299* .