A. Appendix

A.1. Detailed Dataset Description

CREMA-D (Creole Multimodal Affect Database) CREMA-D is a multimodal dataset designed for emotion recognition research. It contains audio and video recordings of actors from the Haitian Creole culture portraying various emotional states. The dataset is valuable for studying the cross-cultural aspects of emotion recognition and includes a range of emotions expressed through speech and facial expressions.

Kinetic-Sound Kinetic-Sound is a dataset that combines the fields of computer vision and audio processing. It consists of synchronized video and audio recordings of everyday objects and actions, providing a rich resource for research on audio-visual scene understanding and object recognition. Researchers use Kinetic-Sound to explore how audio and visual information can complement each other for better recognition and understanding of real-world scenes.

Food-101 Food-101 is a widely-used dataset for food image classification and recognition. It contains over 100,000 images of 101 different food categories, making it a valuable resource for training and evaluating machine learning models for food recognition tasks. Researchers and developers use Food-101 to build applications for automated food identification, dietary analysis, and more.

MVSA (Multimodal Visual Sentiment Analysis) : MVSA is a dataset designed to study sentiment analysis in multimedia content. It combines text, image, and audio data to capture the sentiment expressed in social media posts. Researchers use MVSA to develop advanced models for understanding the emotions and sentiments conveyed in multimedia content, which is crucial for applications like social media monitoring and sentiment analysis in marketing.

IEMOCAP (Interactive Emotional Dyadic Motion Capture) IEMOCAP is a unique dataset created for research in emotion recognition and speech processing. It consists of audio and motion-capture data recorded during natural, emotionally charged conversations between actors. IEMOCAP is widely used to advance the understanding of emotion in speech and gesture, as well as for developing emotion-aware conversational AI systems and therapy applications.

These datasets play crucial roles in advancing research in their respective fields, providing valuable resources for developing and evaluating models and algorithms related to emotion recognition, audio-visual scene understanding, food recognition, sentiment analysis, and emotion in speech and gesture.

A.2. Details of Baselines

In this section, we provide a comprehensive overview of the baseline methods employed in our multimodal learning framework.

Summation Summation fusion, also known as element-wise addition, serves as a fundamental technique for integrating information from diverse modalities. It combines representations from different modalities through a straightforward element-wise summation process. While this method effectively merges information from multiple sources, it does so without explicitly modeling interactions between modalities. In the summation method, each modality-specific encoder is equipped with a fully connected layer. The input and output dimensions of this layer correspond to the output dimension of the encoder and the number of classes, respectively. Subsequently, the unimodal outputs from these fully connected layers are summed to obtain a fusion output. The fusion output is then utilized to calculate the loss and update all parameters of the modality-specific encoders and fully connected layers.

Concatenation Concatenation fusion involves the concatenation of feature vectors from different modalities along a specific axis. This technique enables the model to perceive combined information but does not inherently capture cross-modal interactions. It serves as a foundational approach for feeding multimodal data into neural networks. In the concatenation method, a single fully connected layer is employed. The input dimension of this layer equals the sum of the output dimensions of all encoders, while the output dimension corresponds to the number of classes. During forward propagation, all encoder outputs are concatenated and passed into the fully connected layer to obtain the fusion output. During backpropagation, the loss computed using the fusion output and ground truth is used to update all parameters, including those of the encoders and the fully connected layer.

Late fusion [14] Late fusion approaches involve processing each modality separately and subsequently combining the results at a later stage. This method provides flexibility in handling modalities independently before fusing them, but may miss capturing early interactions between data sources. In the late fusion method, multiple fully connected layers correspond to each modality-specific encoder, similar to the summation approach. However, in this method, there are no cross-modal interactions. All encoders and their corresponding fully connected layers are trained independently on a single modality. The fusion output is calculated as the average of all unimodal outputs.

FiLM [27] FiLM is an advanced multimodal fusion technique that introduces condition encoding and feature modulation. By modulating visual features based on textual conditions, FiLM enables dynamic adjustments of visual representations in response to textual inputs. This method is particularly effective in tasks requiring fine-grained alignment between modalities. In the FiLM method, either audio or text serves as the condition encoder's input to conditionally encode the remaining modalities. The condition encoder is implemented as a fully connected layer, with its input dimension matching the output dimension of the modality-specific encoder. The output dimension of the condition encoder is twice its input dimension, and the encoding is split into γ and β components, which are used to modify the output of other modalities. The modified output is then appended to a fully connected layer for classification and loss computation during optimization.

BiGated [19] BiLinear Gated Fusion leverages bilinear pooling and gating mechanisms to capture intricate interactions between modalities, explicitly modeling cross-modal correlations. This approach provides a more expressive fusion strategy for capturing nuanced relationships between different data sources. In the BiGated method, multiple fully connected layers correspond to each encoder, similar to the summation approach. Additionally, a fusion layer aggregates the outputs from all modalities, as in the concatenation method. However, one modality's hidden state undergoes an activation function (sigmoid in our experiments) to derive a gated weight. This weight is used to modulate the other hidden states before they are appended to the fully connected layer to obtain the fusion output.

OGM-GE [26] OGM-GE addresses under-optimization for specific modalities in multimodal learning by modifying the gradients' values. This approach balances attention weights for each modality, reducing the impact of less informative modalities on the fusion output during training. In the OGM-GE method, which is implied in the setting of concatenation, modality-specific coefficients κ_t^u are calculated according to the algorithm outlined in the original paper. These coefficients are then used to scale all gradients during backpropagation, balancing the optimization of each modality.

QMF [48] Quality-aware Multimodal Fusion (QMF) is designed to efficiently fuse various modalities, particularly in information imbalance situations. Similar to late fusion, QMF introduces a multimodal loss to learn cross-modal information and a regularization term to encourage modalities to pay more attention to challenging samples. In our experiments, we applied QMF similarly to the late fusion setting. The loss calculation aligns with the methodology described in the QMF papers.

A.3. Details of experimental setup

A.3.1 Learning with complete modality

To evaluate the effectiveness of our method across diverse models and datasets, we employed three distinct encoders:

ResNet-18 Based Network We utilized a ResNet-18 based network for the CREMA-D and KS datasets. The ResNet-18 architecture is a member of the ResNet family, specifically designed to address the challenge of vanishing gradients during the training of deep neural networks. It introduces the concept of residual connections, also known as skip connections or shortcut connections, which facilitate the direct flow of information across network layers. This alleviates the vanishing gradient issue, enabling the training of exceptionally deep networks. In our experiments, we initialized the weights of ResNet-18 using the standard initialization method.

M3AE (Multimodal Multimodal Contrastive Learning Based Encoder) For the Food-101 and MVSA datasets, we employed the M3AE encoder. M3AE is a large-pretrained model designed for both vision and language data, leveraging multimodal contrastive learning. It has demonstrated outstanding performance in various downstream tasks. We initiated the M3AE encoder by loading its base model, which has been released publicly.

M3AE + CAVMAE In the case of the IEMOCAP dataset, we adopted a combination of encoders. The acoustic encoder was based on CAVMAE (Audio-Vision Task Pretrained Encoder), a large-pretrained model specialized for audio-vision tasks. For the visual-textual modality, we utilized M3AE. The weight initialization for the CAVMAE encoder was performed by loading the pre-trained cavmae-audio model.

In all experiments, we employed a shared head consisting of a fully connected layer. The input dimension for this layer was set to 512 for experiments with the base model and 768 for experiments with the large-pretrained model. We employed the Stochastic Gradient Descent (SGD) optimizer for all experiments with a momentum value of 0.9. The initial learning rate was set to 0.001. Learning rate decay was applied at regular intervals during training, with a decay ratio of 0.1. In all experiments, we utilized a batch size of 64.

The choice of features varied across datasets and modalities: For the CREMA-D and KS datasets, we utilized the fbank acoustic feature, while the visual feature consisted of a concatenation of three transformed images. In the case of the Food-101, MVSA, and IEMOCAP datasets: Textual Feature:

We employed tokenized text extracted using a BERT-based model. Visual Feature: For these datasets, the visual feature was based on transformed images. Acoustic Feature (IEMOCAP only): For the IEMOCAP dataset, the acoustic feature was incorporated into the CAVMAE encoder and consisted of the fbank feature.

A.3.2 Learning with missing modalities

In our experiments focusing on missing modalities within the IEMOCAP dataset, we employed a diverse set of feature extraction models to capture textual, visual, and acoustic information. Here, we provide detailed descriptions of the feature extraction models and their respective advantages:

We extracted textual features using BERT (Bidirectional Encoder Representations from Transformers), a groundbreaking natural language processing model developed by Google AI in 2018. BERT revolutionized the field of NLP by introducing a pre-trained model capable of understanding contextual relationships in both directions (left-to-right and right-to-left) within a sentence. This bidirectional approach enables BERT to capture intricate word relationships, making it highly effective for a wide range of NLP tasks, including text classification, question-answering, and sentiment analysis.

For visual feature extraction, we relied on MANet, an advanced pre-trained model specializing in visual tasks. MANet builds upon the success of convolutional neural networks (CNNs) by incorporating multi-attention mechanisms. This unique design allows MANet to efficiently process and understand visual information by selectively attending to relevant regions within an image. Consequently, MANet excels in tasks such as image recognition, object detection, and scene understanding, enabling it to capture detailed context and intricate visual relationships.

To extract acoustic features, we utilized Wav2vec, an innovative pre-trained model developed by Facebook AI in 2019. Wav2vec is specifically designed for speech and audio processing tasks, offering the capability to directly convert raw audio waveforms into meaningful vector representations. This model has significantly enhanced the accuracy and efficiency of various audio-related applications, including automatic speech recognition, voice activity detection, and audio classification.

For all methods, we employed modality-specific encoders consisting of a sequence of three fully connected layers. The input dimensions for the acoustic, visual, and textual modalities were set to 512, 1024, and 1024, respectively. The embedding size for all fully connected layers was fixed at 128. The classifier utilized in all methods was a fully connected layer. Both the input and output dimensions of the classifier were set to 128 and corresponded to the number of classes specific to the dataset. The batch size for all experiments was set to 64. We initiated training with an initial learning rate of 0.001, which was reduced in every iteration with a decay ratio of 0.1.

This comprehensive setup allowed us to explore the impact of missing modalities in the IEMOCAP dataset while leveraging state-of-the-art feature extraction models for textual, visual, and acoustic data. The combination of BERT, MANet, and Wav2vec, along with carefully tuned network architectures and training parameters, enabled us to conduct rigorous and insightful experiments in multimodal learning.

A.4. Full Results

A.4.1 Full result of complete multimodal learning

A.4.2 Full Results of Ablation Studies

In Table 6, we report the full results of ablation studies with 95% standard deviation.

A.4.3 Full Results of CLIP

Table 5. The full results on audio-video (A-V), image-text (I-T), and audio-image-text (A-I-T) datasets. Both the results of only using a single modality and the results of combining all modalities ("Multi") are listed. We report the average test accuracy (%) of three random seeds. The best results and second best results are **bold** and underlined, respectively.

Type	Data		Sum	Concat	Late Fusion	FiLM	BiGated	OGM-GE	QMF	**MLA (Ours)**
A-V	CREMA-D	Audio	54.14 ± 0.92	55.65 ± 0.82	52.17 ± 1.12	53.89 ± 0.75	51.49 ± 1.28	53.76 ± 0.98	$\mathbf{59.41} \pm 0.71$	59.27 ± 1.23
		Video	18.45 ± 1.07	18.68 ± 0.76	55.48 ± 0.71	18.67 ± 1.21	17.34 ± 1.11	28.09 ± 1.17	39.11 ± 1.03	$\mathbf{64.91} \pm 1.10$
		Multi	60.32 ± 0.78	61.56 ± 0.91	66.32 ± 1.08	60.07 ± 1.25	59.21 ± 1.14	68.14 ± 0.79	63.71 ± 1.12	$\mathbf{79.70} \pm 0.87$
	KS	Audio	48.77 ± 0.95	49.18 ± 0.76	47.87 ± 1.10	48.67 ± 0.83	49.96 ± 1.21	48.87 ± 1.05	51.57 ± 1.03	$\mathbf{54.67} \pm 0.92$
		Video	24.53 ± 1.12	24.67 ± 1.07	46.76 ± 0.85	23.15 ± 0.98	23.77 ± 1.14	29.73 ± 1.06	32.19 ± 0.92	$\mathbf{51.03} \pm 1.09$
		Multi	64.72 ± 0.97	64.84 ± 1.05	65.53 ± 0.89	63.33 ± 0.76	63.72 ± 1.13	65.74 ± 1.08	65.78 ± 0.97	$\mathbf{71.35} \pm 1.22$
I-T	Food-101	Image	4.57 ± 0.88	3.51 ± 1.22	58.46 ± 1.03	4.68 ± 0.97	14.20 ± 1.18	22.35 ± 1.27	45.74 ± 1.09	$\mathbf{69.60} \pm 0.89$
		Text	85.63 ± 1.14	86.02 ± 1.05	85.19 ± 1.21	85.84 ± 0.92	85.79 ± 1.10	85.17 ± 1.09	84.13 ± 1.27	$\mathbf{86.47} \pm 0.86$
		Multi	86.19 ± 0.86	86.32 ± 1.18	90.21 ± 1.02	87.21 ± 1.05	88.87 ± 0.88	87.54 ± 1.03	92.87 ± 1.01	$\mathbf{93.33} \pm 0.92$
	MVSA	Text	73.33 ± 1.03	75.22 ± 0.92	72.15 ± 1.07	74.85 ± 0.98	73.13 ± 1.08	74.76 ± 0.87	74.87 ± 1.15	$\mathbf{75.72} \pm 0.79$
		Image	28.46 ± 1.12	27.32 ± 1.09	45.24 ± 0.97	27.12 ± 1.18	28.15 ± 0.88	31.98 ± 1.03	32.99 ± 1.09	$\mathbf{54.99} \pm 1.12$
		Multi	76.19 ± 1.01	76.25 ± 0.95	76.88 ± 0.82	75.34 ± 1.07	75.94 ± 0.88	76.37 ± 0.98	77.96 ± 1.06	$\mathbf{79.94} \pm 0.97$
A-I-T	IEMOCAP	Audio	39.79 ± 1.08	41.93 ± 0.89	43.12 ± 0.97	41.64 ± 1.06	42.23 ± 0.88	41.38 ± 1.13	42.98 ± 0.95	$\mathbf{46.29} \pm 1.02$
		Image	29.44 ± 0.97	30.00 ± 0.88	32.38 ± 0.92	29.85 ± 1.06	27.45 ± 1.08	30.24 ± 1.02	31.22 ± 1.07	$\mathbf{37.63} \pm 0.78$
		Text	65.16 ± 0.88	67.84 ± 0.97	68.79 ± 1.07	66.37 ± 0.95	65.16 ± 1.08	70.79 ± 1.03	75.03 ± 0.79	$\mathbf{73.22} \pm 1.09$
		Multi	74.18 ± 1.03	75.91 ± 0.92	74.96 ± 0.97	74.32 ± 1.12	73.34 ± 1.05	76.17 ± 1.01	76.17 ± 0.95	$\mathbf{78.92} \pm 1.07$

Table 6. We report the test accuracy percentages (%) on the IEMOCAP dataset using three different seeds, while applying varying modality missing rates to audio, image, and text data. The best results are highlighted in **bold**, while the second-best results are underlined.

Method	Modality Missing Rate (%)						
	10	20	30	40	50	60	70
Late Fusion	72.95 ± 1.12	69.06 ± 0.88	64.89 ± 1.25	61.09 ± 0.99	56.48 ± 1.18	52.41 ± 1.01	45.07 ± 1.21
QMF	73.49 ± 1.10	71.33 ± 1.27	65.89 ± 0.78	62.27 ± 0.95	57.94 ± 1.07	55.60 ± 0.84	50.25 ± 1.09
CCA	65.19 ± 0.95	62.60 ± 1.19	59.35 ± 0.88	55.25 ± 1.23	51.38 ± 1.05	45.73 ± 1.27	30.61 ± 1.10
DCCA	57.25 ± 1.28	51.74 ± 1.01	42.53 ± 1.29	36.54 ± 1.18	34.82 ± 0.79	33.65 ± 1.07	41.09 ± 1.24
DCCAE	61.66 ± 1.13	57.67 ± 1.22	54.95 ± 1.06	51.08 ± 1.01	45.71 ± 0.87	39.07 ± 0.98	41.42 ± 1.28
AE	71.36 ± 0.94	67.40 ± 0.79	62.02 ± 1.18	57.24 ± 0.94	50.56 ± 0.77	43.04 ± 0.96	39.86 ± 0.82
CRA	71.28 ± 1.19	67.34 ± 1.04	62.24 ± 1.03	57.04 ± 0.92	49.86 ± 1.15	43.22 ± 0.99	38.56 ± 1.22
MMIN	71.84 ± 0.97	69.36 ± 1.05	66.34 ± 1.11	63.30 ± 1.10	60.54 ± 1.17	57.52 ± 1.13	55.44 ± 1.06
IF-MMIN	71.32 ± 0.78	68.29 ± 0.98	64.17 ± 1.01	60.13 ± 1.19	57.45 ± 1.12	53.26 ± 1.22	52.04 ± 0.91
CPM-Net	55.29 ± 0.83	53.65 ± 0.96	52.52 ± 1.29	51.01 ± 0.94	49.09 ± 1.04	47.38 ± 1.25	44.76 ± 0.88
TATE	67.84 ± 1.02	63.22 ± 1.08	62.19 ± 0.98	60.36 ± 1.06	58.74 ± 1.21	57.99 ± 1.03	54.35 ± 1.07
MLA (Ours)	$\mathbf{75.07} \pm 1.04$	$\mathbf{72.33} \pm 1.16$	$\mathbf{68.47} \pm 1.22$	$\mathbf{67.00} \pm 0.98$	$\mathbf{63.48} \pm 0.87$	$\mathbf{59.17} \pm 0.92$	$\mathbf{55.89} \pm 1.03$

Table 7. Results on the Food-101 dataset achieved by changing the encoders to the CLIP pre-trained model. We report the average test accuracy (%) from three different seeds.

Method	Food-101		
	Image	Text	Multi
CLIP	63.07 ± 0.12	83.98 ± 0.08	93.07 ± 0.08
CLIP + MLA (Ours)	$\mathbf{72.22 \pm 0.10}$	$\mathbf{85.34 \pm 0.06}$	$\mathbf{93.47 \pm 0.04}$