Multimodal Representation Learning by Alternating Unimodal Adaptation

Xiaohui Zhang,* Jaehong Yoon, Mohit Bansal, Huaxiu Yao
UNC-Chapel Hill
xzhang47@nd.edu, huaxiu@cs.unc.edu

Abstract

Multimodal learning, which integrates data from diverse sensory modes, plays a pivotal role in artificial intelligence. However, existing multimodal learning methods often struggle with challenges where some modalities appear more dominant than others during multimodal learning, resulting in suboptimal performance. To address this challenge, we propose MLA (Multimodal Learning with Alternating Unimodal Adaptation). MLA reframes the conventional joint multimodal learning process by transforming it into an alternating unimodal learning process, thereby minimizing interference between modalities. Simultaneously, it captures cross-modal interactions through a shared head, which undergoes continuous optimization across different modalities. This optimization process is controlled by a gradient modification mechanism to prevent the shared head from losing previously acquired information. During the inference phase, MLA utilizes a test-time uncertainty-based model fusion mechanism to integrate multimodal information. Extensive experiments are conducted on five diverse datasets, encompassing scenarios with complete modalities and scenarios with missing modalities. These experiments demonstrate the superiority of MLA over competing prior approaches. Our code is available at https://github.com/Cecile-hi/MLA.

1. Introduction

Multimodal learning, which draws inspiration from the multi-sensory perception mechanisms in humans, has gained significant prominence in the field of artificial intelligence [31, 32, 42]. However, recent multimodal learning methods often struggle to fully integrate rich multimodal knowledge across different modalities, and we argue that a key factor is *modality laziness*. In multimodal representation learning, some modalities are more dominant than others [9, 26], so the model will optimize for these dominant modalities and tend to ignore others, resulting in suboptimal performance [17, 30, 37]. This is because collected multimodal data are often not well entangled with each other, or their

data size varies. In a more extreme scenario, critical modality data may be missing depending on the conditions during the data collection phase [20]. This is particularly one of the main challenges in multimodal learning on uncurated real-world data.

A few recent works have been introduced to balance the influence of dominating versus subordinate modalities in the optimization process [26, 48]. However, these methods necessitate joint optimization of different modes to update the multiple modality-specific encoders simultaneously, which degenerates the adaptation for subordinate modalities to some extent, thereby limiting overall multi-modal performance [37]. In contrast, we aim to tackle this problem in a conceptually different way by decomposing the conventional multimodal joint optimization scenario into an alternating unimodal learning scenario, leading to an approach named **M**ultimodal **L**earning with **A**lternating Unimodal Adaptation (**MLA**). The key idea of MLA is to alternately optimize the encoder of each modality, while simultaneously integrating cross-modal information.

Concretely, as shown in Figure 1, the predictive function of each modality in our approach includes a modality-specific encoder and a shared head across all modalities. In the alternating unimodal learning paradigm, the predictive functions for each modality are optimized alternately to eliminate interference across modalities. Simultaneously, the shared head is optimized continuously across modalities, essentially capturing cross-modal information. However, in this optimization process, the head is susceptible to losing previously learned information from other modalities when it encounters a new modality, which is referred to as modality forgetting. To address this issue, we introduce a gradient modification mechanism for the shared head to encourage the orthogonalization of gradient directions between modalities. After learning the modality-specific encoders and the shared head, we further propose a test-time dynamic modality fusion mechanism to integrate multimodal information. Since there are information gaps among different modalities contributing to the prediction, we evaluate the significance of each modality and use this evaluation to assign weights to the predictions generated by each modality. Our method

---

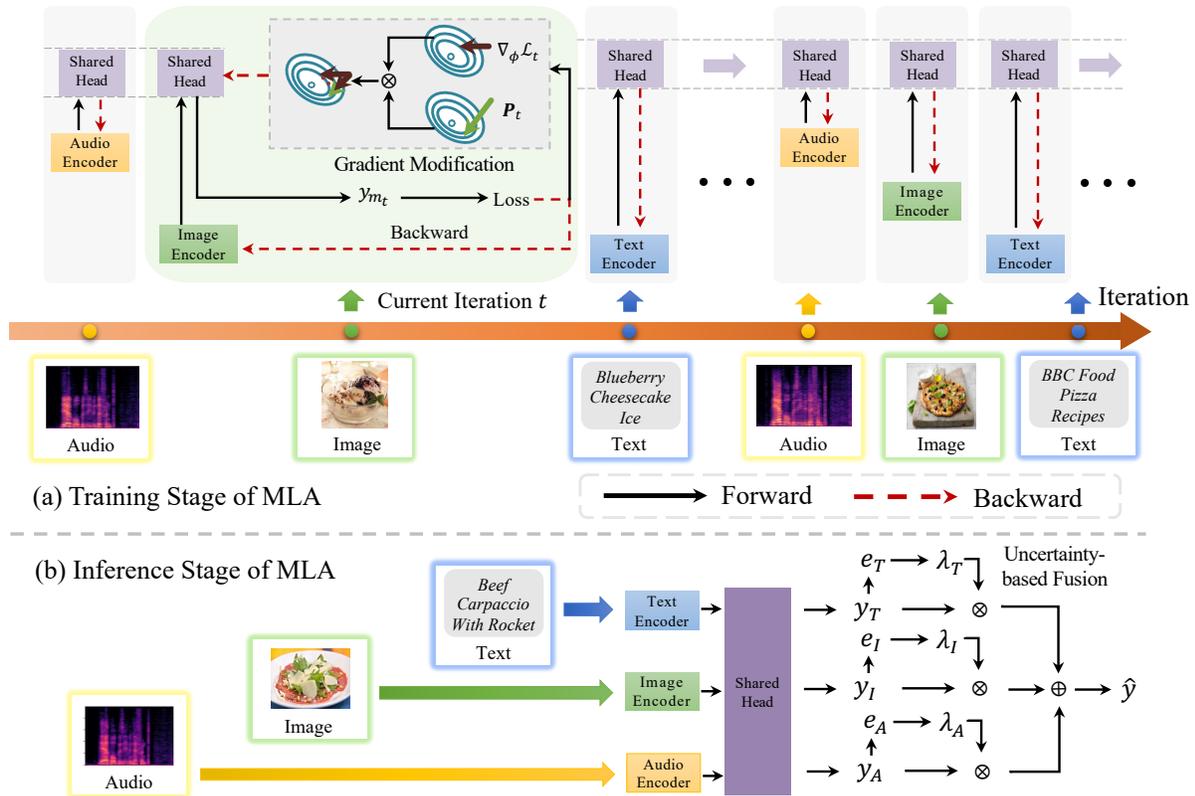*Work was done during Xiaohui Zhang's remote internship at UNC.

Figure 1. The framework of MLA: (a) Training stage, where we employ an alternating unimodal adaptation process, processing only one modality at each iteration. The shared head captures cross-modal information, and we also introduce gradient modification to prevent forgetting learned modality information from previous iterations; (b) Testing stage, where we introduce an uncertainty-based test-time multimodal fusion approach to combine multimodal information.

for gauging the importance of each modality relies on measuring the level of uncertainty observed in the predictions associated with that modality. This mechanism is motivated by the hypothesis that when one modality exhibits higher uncertainty in its predictions, it is more prone to producing incorrect predictions.

Our primary contribution of this paper is MLA, which introduces an innovative alternating unimodal optimization approach in multimodal learning. This approach not only enables relatively independent optimization within each modality but also preserves cross-modal interactions. MLA is also compatible with scenarios involving both complete and missing modalities in the learning process. In both scenarios, our empirical results demonstrate the promise of MLA in addressing modality laziness and enhancing multimodal learning performance compared to the best prior methods. Furthermore, we show that MLA can enlarge the modality gap, offering further insights into the performance improvements it achieves.

2. Related Work

Imbalanced Modality Contributions to Multimodal Learning Leveraging multiple modalities that contain unique yet complementary representations is crucial for understanding and solving the real world problems [1, 7, 8, 21, 24, 34, 40, 54–56]. However, training on multimodal data simultaneously is challenging since different modalities have significant discrepancies in their data distributions, learning architectures, and target tasks. These varying characteristics in multiple modalities hinder the model from integrating the knowledge from different senses and also lead to the problem of *modality laziness* [10, 30, 37, 41], a phenomenon where some modalities appear more dominant than others during multimodal learning. To tackle this critical problem and utilize rich information from subordinate modalities, several approaches have recently been suggested [11, 16, 18, 38]. Representatively, OGM-GE [26] proposes a balanced multimodal learning method that corrects the contribution imbalance of different modalities by encouraging intensive gradient updating from suboptimal modalities. QMF [48] introduces a robust multimodal learning method that mitigates the impact of low-quality or noisy modalities by estimating

the energy-based score of each modality.

While these methods mitigated imbalanced modality contributions to multimodal learning, they require simultaneous optimization of multiple modules for different sensory data by design, which can result in undesired interference in training on different modalities. Instead, our proposed method can avoid this inherent problem by rethinking multimodal learning to alternating unimodal learning, where the model can independently learn the corresponding unimodal information at each training step and capture the abundant cross-modal information. This paradigm allows the model to optimize each modality sufficiently, leading to building a balanced and effective multimodal model.

Learning with Missing Modalities Some modality data is often missing in real-world scenarios for various reasons, such as cost, accidents, or privacy concerns. Such missing modality issue [15, 22, 23, 45, 52] is also a crucial challenge for balanced training on multimodal data, which can be regarded as an extreme case of modality laziness. To resolve this problem, one conventional approach is a data imputation strategy [33, 53], which generates estimated data/embedding based on statistics from existing multimodal data pairs. This approach often works well on small data, but is often inappropriate for large data or when large chunks of data are missing. Alternatively, several works have addressed the problem of missing modality without the need for imputation [22, 23, 46]. [46] introduces a new framework for partial multi-view learning, where each incoming data is associated with only one of the multiple views (i.e., modalities). They adopt a generative adversarial strategy to estimate missing data and improve the quality of multi-view representation. SMIL [22] addresses the problem of multimodal learning with severely missing modalities (e.g., 90%). It utilizes a Bayesian meta-learning framework that learns to perturb the latent feature space and then estimates the representation of missing modalities from accessible sensory data. However, such previous works are tailored to mitigate the problem of missing modality in multimodal learning, resulting in their limited generality on diverse multimodal problems. Unlike previous works, our proposed MLA is broadly applicable to traditional multimodal learning tasks as well as challenging scenarios with missing modality data, and we have demonstrated its effectiveness by outperforming solid baselines on these tasks.

3. Multimodal Learning with Alternating Unimodal Adaptation

This section presents our proposed method to address the modality laziness issue, named *Multimodal Learning with Alternative Unimodal Adaptation* (**MLA**). Motivated by the challenge of information imbalance in different modalities,

MLA aims to reframe conventional joint training scheme for multimodal data into the context of a alternating unimodal learning framework, as illustrated in Figure 1. Specifically, during the training phase (Figure 1(a)), our approach alternately learns the encoders for each modality while simultaneously maintaining cross-modal information using a shared head. We introduce a gradient modification mechanism to prevent the shared head from forgetting previously learned modality information. During inference, our approach dynamically fuses multimodal information by evaluating the uncertainty of the prediction for each modality (Figure 1(b)). Next, we will introduce the three key stages: *alternating unimodal learning*, *learning cross-modal information without modality forgetting*, and *test-time dynamic modality fusion*.

3.1. Alternating Unimodal Learning

In multimodal learning scenarios, the phenomenon known as modality laziness arises due to the inadequate optimization of less informative modalities when these are learned in conjunction with others, subsequently leading to suboptimal fusion performance. To tackle this problem, we propose an alternating unimodal learning paradigm. Here, each modality undergoes an independent optimization process, eliminating interference across modalities, thus allowing each modality to reach its full potential without being overshadowed by more informative ones.

Let us consider solving multimodal learning problem with M modalities, where the dataset for the m-th modality is denoted as $\mathcal{D}_m = (X_m, Y_m) = \{(x_{m,k}, y_{m,k})\}_{k=1}^{N_m}$. Here, N_m represents the number of examples in m, and each modality is associated with a predictive function formulated as $f_m = g \circ h_m$, wherein the function h_m serves as the modality-specific encoder and the function g denotes the shared head across all modalities. Given T total training steps, the model receives data exclusively from one modality. We determine the input modality m_t at the timestep t as follows:

$$m_t = t \bmod M, \quad \text{where} \quad t < M. \tag{1}$$

At each training step t, we iteratively optimize the multimodal model by minimizing the predictive risk $\mathcal{L}_t$ for training examples within the corresponding single modality data $\mathcal{D}_{m_t}$:

$$\begin{aligned}
\mathcal{L}_t &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{m_t}}[\ell(f_{m_t}(x; \theta_{m_t}, \phi), y)] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}_{m_t}}[\ell(g(h_{m_t}(x; \theta_{m_t}); \phi), y)].
\end{aligned} \tag{2}$$

where θ_{m_t} and ϕ are learnable parameters of encoder h_{m_t} and the shared head g, respectively. This allows MLA to capture a rich representation of all available modalities while avoiding the multimodal model from learning only dominant modality information (i.e., modality laziness). It's also worth noting that MLA does not require paired multimodal

data during the training phase, making it a natural fit for scenarios with extreme modality laziness, such as learning with missing modalities.

3.2. Learning Cross-Modal Information without Modality Forgetting

In multimodal learning, besides isolating the optimization process for each modality to prevent modality laziness, capturing cross-modal interaction information stands out as another crucial aspect. Multimodal learning with the isolated optimization process for each modality surprisingly mitigates the problem of modality laziness, but there is a caveat. Beyond retaining multiple modality representations, the desired multimodal model should be able to capture interactions between modalities. In equation 2, our framework adopts the shared head g across all given modalities, enabling the capture of cross-modal interaction information throughout the process. Nonetheless, this sequential optimization process can pose a new challenge: the head g is prone to forgetting the information of previously trained modalities when learning new ones, a phenomenon termed as *modality forgetting*. This issue can significantly undermine the effectiveness of cross-modal information learning.

Inspired by weight modification methods [49–51], we address modality forgetting by introducing a gradient modification matrix $\mathbf{P}_t$ at each iteration t to rectify the gradients of parameter ϕ of the shared head g before starting to learn new modality. This gradient modification ensures that the parameter update direction is orthogonal to the plane spanned by the encoded feature from the previous modality. Thus, we ensure that applying the gradients to one modality minimally interferes with its prior modality. Specifically, the optimization process of parameters ϕ of the shared head g during iteration t is defined as follows:

$$\phi_t = \phi_{t-1} - \gamma \begin{cases} \nabla_\phi \mathcal{L}_t & \text{if } t = 0, \\ \mathbf{P}_t \nabla_\phi \mathcal{L}_t & \text{if } t > 0, \end{cases} \quad (3)$$

where $\mathcal{L}_t$ is defined in equation 2.

To obtain the gradient modification matrix $\mathbf{P}_t$, we leverage Recursive Least Square algorithm [29]. Specifically, we define the average output from the encoder as:

$$\overline{h_{m_t}(x)} = \frac{1}{N_{m_t}} \sum_{k=1}^{N_{m_t}} h_{m_t}(x_{m_t,k}). \quad (4)$$

Suppose s represents the dimension of the output from the encoder. At each iteration t, the corresponding modification matrix $\mathbf{P}_t \in \mathbb{R}^{s \times s}$ is then obtained as follows:

$$\mathbf{P}_t = \mathbf{P}_{t-1} - \mathbf{q}_t \left[\overline{h_{m_t}(x)} \right]^T \mathbf{P}_{t-1},$$

$$\text{where } \mathbf{q}_t = \frac{\mathbf{P}_{t-1} \overline{h_{m_t}(x)}}{\alpha + \left[\overline{h_{m_t}(x)} \right]^T \mathbf{P}_{t-1} \overline{h_{m_t}(x)}}, \quad (5)$$

Algorithm 1: Multimodal Learning with Alternating Unimodal Adaptation (**MLA**)

Input: Multimodal datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M$, # of training iterations T

1 **for** $t = 1$ **to** T **do**
2   Calculate the modality assignment m_t at iteration t as equation 1;
3   Compute the loss $\mathcal{L}_t$ following equation 2 with m_t;
4   Update the modality-specific parameters θ_{m_t} with gradient descent;
5   Compute the gradient modification matrix $\mathbf{P}_t$ as described in equation 5;
6   Update the shared parameters ϕ_t using equation 3;

7 **Inference stage:**
8 **for** *Test example (x_r, y_r) involving M modalities* **do**
9   Calculate the prediction uncertainty $e_{m,r}$ for each modality m following equation 7;
10   Determine the modality importance coefficients $\lambda_{m,r}$ using equation 8;
11   Combine the predictions from all modalities as equation 6 and get the predicted value $\hat{y}_r$;

---

where α is a predefined hyperparameter used to prevent denominator from being zero. The modification matrix is initialized as an identity matrix before training. By introducing the gradient orthogonalization process to calibrate the weight update direction for the shared head g, we mitigate the interference across consecutive modalities and facilitate a more effective capture of cross-modal information.

3.3. Test-Time Dynamic Modality Fusion

After learning the modality-specific encoders and the shared head during the training process, we focus on how to effectively integrate multimodal information to make prediction during inference time. To achieve this multimodal fusion, we employ a weighted combination of predictions from each modality. Specifically, for a given test example (x_r, y_r), which involves M modalities, we calculate its prediction as follows:

$$\hat{y}_r = \sum_{m=1}^{M} \lambda_{m,r} f_m(x_{m,r}; \theta_m^*, \phi^*), \quad (6)$$

where $\lambda_{m,r}$ signifies the importance of modality m in predicting the labels for the test example r. The parameters θ_m^* and ϕ^* are optimized values associated with the encoder h_m and the shared head g, respectively.

To determine the value of modality importance coefficient λ_m, MLA performs under the hypothesis that when

one modality exhibits higher uncertainty in its predictions, it becomes more prone to making incorrect predictions. Consequently, we leverage prediction uncertainty as a proxy to gauge the importance of each modality. It is worth noting that every modality can reflect strong uncertainty, regardless of whether it is dominating or subordinate modalities. Specifically, we begin by assessing the uncertainty $e_{m,r}$ using entropy of each individual modality's output as follows:

$$e_{m,r} = -p_{m,r}^T \log p_{m,r},$$
$$\text{where} \ \ p_{m,r} = \text{Softmax}(f_m(x_{m,r}; \theta_m^*, \phi^*)). \tag{7}$$

Here, $\text{Softmax}(\cdot)$ converts the output logits to the probability $p_{m,r}$. A higher entropy $e_{m,r}$ indicates lower confidence in the prediction, leading to a smaller importance weight during the fusion process. Based on this, we calculate the importance weight for a modality m as:

$$\lambda_{m,r} = \frac{\exp \left(\max\limits_{m=1,\dots,M} e_{m,r} - e_{m,r} \right)}{\sum_{v=1}^{M} \exp \left(\max\limits_{m=1,\dots,M} e_{m,r} - e_{v,r} \right)}. \tag{8}$$

By introducing the test-time dynamic fusion mechanism that explicitly considers the predictive uncertainty associated with each modality, MLA is better equipped to handle scenarios with imbalances in modality-specific information, enhancing the effectiveness of multimodal fusion. The whole training and inference pipeline is provided in Algorithm 1.

4. Experiments

In this section, we evaluate the performance of MLA, aiming to answer the following questions: **Q1:** Compared to prior approaches, can MLA overcome modality laziness and improve multimodal learning performance? **Q2:** How does MLA perform when faced the challenge of missing modalities? **Q3:** Can the proposed modules (e.g., test-time dynamic fusion) effectively improve performance? **Q4:** How does MLA change the modality gap in multimodal learning?

4.1. Learning with complete modalities

4.1.1 Experimental Setups

Datasets. We utilize a set of five datasets with different tasks to assess the performance of learning with complete modalities (see Appendix A.1 for detailed data statistics and descriptions). **(1)&(2) CREMA-D** [6] **and Kinetic-Sound (KS)** [3] belong to the category of *audio-video* datasets. CREMA-D provides audio and video recordings depicting various emotions within the Haitian Creole culture, while KS combines video and audio data for object and action recognition. **(3)&(4) Food-101** [39] **and MVSA** [25] are both *image-text* datasets. Food-101 comprises over 100,000 food

images accompanied by corresponding texts, with a focus on food classification. MVSA concentrates on sentiment classification in multimedia posts through the utilization of both text and image data. **(5) IEMOCAP** [5] is a *audio-image-text* dataset that captures emotions across audio, vision, and text data during natural conversations.

Baselines. We conducted a comprehensive comparison of MLA with (1) *conventional multimodal fusion methods*: including summation (Sum), concatenation (Concat), late fusion [14]; (2) *modulation-based fusion methods*: including FiLM [27], and BiLinear Gated (BiGated) [19]; (3) *methods for addressing modality-laziness*: including OGM-GE [26] and QMF [48]. For more detailed baseline descriptions, please refer to Appendix A.2.

Unimodal and Multimodal Evaluation. For each baseline, we conduct a comprehensive evaluation and report our performance on datasets using all available modalities. Additionally, we assess performance by utilizing only a single modality's data to make predictions. In this scenario, for fusion methods such as later fusion, FiLM, BiGated, QGM-GE, and QMF, we simply deactivate one modality within the multi-modal network. For the Sum method, we follow the approach outlined in [26], where we input features from the evaluated modality into a shared head. Conversely, for the Concat method, as recommended in Peng et al. [26], we partition the head into multiple subheads and employ the corresponding subhead to assess performance, which reflects the unimodal performance.

Backbone and Hyperparameter Settings. For the audio-video task (CREMA-D and KS datasets), we employ a ResNet-18-based [26] network as the encoder. In the image-text task (Food-101 and MVSA), we utilize M3AE [12], a large pre-trained multimodal masked auto-encoder, as the encoder. For the audio-image-text task (IEMOCAP), we integrate M3AE with another large pre-trained model named CAVMAE [13] as the encoder. In all experiments, a fully connected layer served as the shared head across different modalities. To ensure a fair comparison, all baselines used the same backbone architectures. We determined all other hyperparameters through cross-validation. Details regarding our experimental setups can be found in Appendix A.3.

4.1.2 Result

In Table 1, we report the results of only using a single modality and the results achieved by combining all modalities. The observations from Table 1 reveal several key insights:

- *Firstly*, most conventional fusion and modulation-based methodologies, with the exception of late fusion, faced challenge of modality laziness. This was evidenced by a noticeable performance disparity on numerous datasets between the superior and inferior modality performances.
- *Second*, the late fusion approach addresses modality lazi-

Table 1. 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. Full results with standard deviation are reported in Appendix A.4. 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	55.65	52.17	53.89	51.49	53.76	**59.41**	<u>59.27</u>
		Video	18.45	18.68	<u>55.48</u>	18.67	17.34	28.09	39.11	**64.91**
		Multi	60.32	61.56	66.32	60.07	59.21	<u>68.14</u>	63.71	**79.70**
	KS	Audio	48.77	49.18	47.87	48.67	49.96	48.87	<u>51.57</u>	**54.67**
		Video	24.53	24.67	<u>46.76</u>	23.15	23.77	29.73	32.19	**51.03**
		Multi	64.72	64.84	65.53	63.33	63.72	65.74	<u>65.78</u>	**71.35**
I-T	Food-101	Image	4.57	3.51	<u>58.46</u>	4.68	14.20	22.35	45.74	**69.60**
		Text	85.63	<u>86.02</u>	85.19	85.84	85.79	85.17	84.13	**86.47**
		Multi	86.19	86.32	90.21	87.21	88.87	87.54	<u>92.87</u>	**93.33**
	MVSA	Text	73.33	<u>75.22</u>	72.15	74.85	73.13	74.76	74.87	**75.72**
		Image	28.46	27.32	<u>45.24</u>	27.12	28.15	31.98	32.99	**54.99**
		Multi	76.19	76.25	76.88	75.34	75.94	76.37	<u>77.96</u>	**79.94**
A-I-T	IEMOCAP	Audio	39.79	41.93	<u>43.12</u>	41.64	42.23	41.38	42.98	**46.29**
		Image	29.44	30.00	<u>32.38</u>	29.85	27.45	30.24	31.22	**37.63**
		Text	65.16	67.84	68.79	66.37	65.16	<u>70.79</u>	75.03	**73.22**
		Multi	74.18	75.91	74.96	74.32	73.34	<u>76.17</u>	<u>76.17</u>	**78.92**

ness by training each modality's encoder exclusively on the corresponding unimodal data. However, while late fusion mitigates modality laziness to some extent, it falls short in delivering satisfactory performance when integrating information from all modalities. This limitation primarily stems from its inability to effectively capture cross-modal information.

- *Third*, both OGM-GE and QMF contribute to reducing modality laziness and enhancing multimodal performance to some degree. However, they do not completely bridge the gap between modalities.

- *Finally*, MLA instead consistently outperforms all other methods across all scenarios. This demonstrates MLA's ability to effectively address modality laziness and enhance multimodal learning performance by fully leveraging the information from each modality and capturing cross-modal knowledge.

4.2. Learning with missing modalities

4.2.1 Experimental Setup

Evaluation Strategy Besides learning with complete modalities, we further evaluate the performance on datasets with missing modalities, which could be regarded as an extreme case of modality laziness. Here, follow [20], we apply percentage-based masks to both the training and testing data within the IEMOCAP dataset. Specifically, we randomly mask each modality of each sample with a probability η. Following [43, 47], we select missing rates from the list $[0.1, 0.2, ...0.7]$, maintaining the same missing rate across the training, validation, and testing phases.

Baselines and Hyperparameter Settings We compare MLA with the following baselines: (1) the strongest and most compatible baselines in the experiments of learning with complete modalities, including late fusion [14] and QMF [48]; (2) methods specifically designed for learning with missing modalities, including CCA [15], DCCA [2], DCCAE [36], AE [4], CRA [33], MMIN [53], IF-MMIN [57], CPM-Net [46], and TATE [44]. All baselines use the same backbone models as used in the experiments of learning with complete modalities. We tune the hyperparameters via cross-validation. Detailed baseline descriptions and experimental setups can be found in Appendix A.3.

4.2.2 Results

We report the results in Table 2, showing the performance under different modality missing rates. We observe that:

- All approaches show performance degradation with the increase of modality missing rate. This is what we expected because utilizing all modality data tends to enhance performance when compared to using only partial modality data, as also corroborated by the results in Table 1, where employing multi-modality information outperforms using only single-modality information.

- MLA consistently outperforms the other baselines across all missing rates, including general methods (Late Fusion and QMF) and methods that are specifically designed for tackling missing modality (e.g., MMIN). These results highlight the effectiveness of MLA in addressing the challenge of modality laziness, even under the extreme case of learning with missing modality. This underscores the ef-

Table 2. 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	69.06	64.89	61.09	56.48	52.41	45.07
QMF	<u>73.49</u>	<u>71.33</u>	65.89	62.27	57.94	55.60	50.25
CCA	65.19	62.60	59.35	55.25	51.38	45.73	30.61
DCCA	57.25	51.74	42.53	36.54	34.82	33.65	41.09
DCCAE	61.66	57.67	54.95	51.08	45.71	39.07	41.42
AE	71.36	67.40	62.02	57.24	50.56	43.04	39.86
CRA	71.28	67.34	62.24	57.04	49.86	43.22	38.56
MMIN	71.84	69.36	<u>66.34</u>	<u>63.30</u>	<u>60.54</u>	57.52	<u>55.44</u>
IF-MMIN	71.32	68.29	64.17	60.13	57.45	53.26	52.04
CPM-Net	55.29	53.65	52.52	51.01	49.09	47.38	44.76
TATE	67.84	63.22	62.19	60.36	58.74	<u>57.99</u>	54.35
MLA (Ours)	**75.07**	**72.33**	**68.47**	**67.00**	**63.48**	**59.17**	**55.89**

fectiveness of alternating unimodal learning and test-time dynamic modality fusion mechanisms.

4.3. Analysis of MLA

In this section, we aim to gain a deeper insight into the performance improvements achieved by MLA. Here, we conduct four analyses. Firstly, we conduct an ablation study, focusing on head gradient modification and test-time modality fusion. Secondly, we explore the isolation of modalities within our alternating unimodal learning framework. Third, we analyze the modality gap. Lastly, we conduct experiments to analyze the robustness of ours over pre-trained models.

4.3.1 Ablation Study

We first conduct ablation studies to demonstrate the effectiveness of using head gradient modification in alternating unimodal learning and test-time dynamic fusion. We report the results in Table 3 and observe that: (1) Modifying head weights under the alternate unimodal learning framework enhances the performance of both using the information from a single modality only and multiple modalities ("Multi"). This improvement is expected, as mitigating the effect of modality forgetting enables better integration of cross-modal information, benefiting both unimodal and multimodal learning processes; (2) Using a test-time dynamic fusion mechanism can significantly advance the multimodal learning process. This result is anticipated, as test-time dynamic fusion accounts for modal predictive uncertainty, which usually correlates with model performance. By using predictive uncertainty to measure the modality importance on test examples, the model can more accurately fuse information from multiple modalities; (3) By integrating both strategies, MLA exhibits the best performance, demonstrating its effectiveness in mul-

Table 3. Results of ablation studies on five datasets. We report the average test accuracy (%) of three random seeds. Full results with standard deviation are reported in Appendix A.4. Note that dynamic fusion is only applied in the multimodal fusion process, which does not affect the performance of using a single modality. HGM: head gradient modification; DF: dynamic fusion.

Data		HGM	DF	HGM	DF	HGM	DF	HGM	DF
		✗	✗	✗	✓	✓	✗	✓	✓
CREMA-D	Audio	52.17		52.17		**59.27**		**59.27**	
	Video	55.48		55.48		**64.91**		**64.91**	
	Multi	66.32		72.79		74.51		**79.70**	
KS	Audio	47.87		47.87		**54.66**		**54.66**	
	Video	46.76		46.76		**51.03**		**51.03**	
	Multi	65.53		66.34		70.72		**71.35**	
Food-101	Text	85.19		85.19		**86.47**		**86.47**	
	Image	58.46		58.46		**69.60**		**69.60**	
	Multi	90.21		91.37		91.72		**93.33**	
MVSA	Text	72.15		72.15		**75.72**		**75.72**	
	Image	45.24		45.24		**54.99**		**54.99**	
	Multi	76.88		77.53		79.59		**79.94**	
IEMOCAP	Audio	43.12		43.12		**46.29**		**46.29**	
	Text	68.79		68.79		**73.22**		**73.22**	
	Image	32.38		32.38		**37.63**		**37.63**	
	Multi	74.96		75.42		77.58		**78.92**	

timodal learning.

4.3.2 Analysis of Modality Isolation in Alternating Unimodal Learning

We investigate how the alternating unimodal learning paradigm isolates the unimodal learning and prevents modality laziness. In this context, we selectively masked either the audio or video modality during training, employing various modality masking ratios. The results are presented in

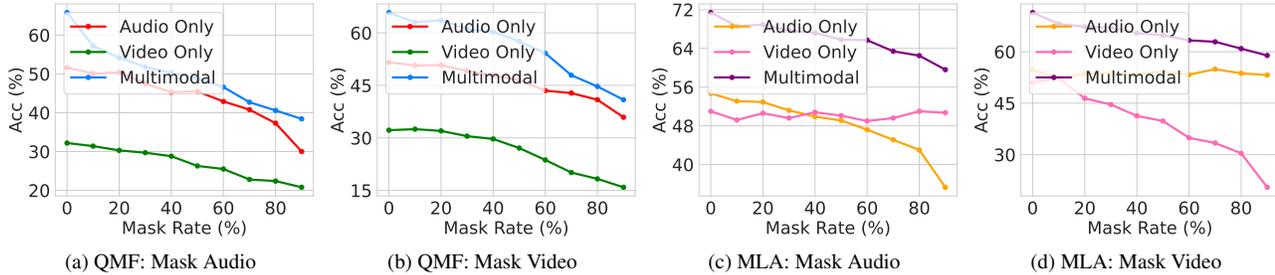| (a) QMF: Mask Audio | (b) QMF: Mask Video | (c) MLA: Mask Audio | (d) MLA: Mask Video |

Figure 2. Visualization of test accuracy (%) on the KS dataset, varying with the ratio of missing modality in audio or video training data.

Figure 2, where the results on QMF are also illustrated for comparison. We observe that the performance of unimodal learning in MLA remains unaffected by the absence of other modalities. In contrast, in QMF, the absence of one modality negatively impacts the performance of another modality. These findings further strengthen our argument that adopting an alternating optimization approach to learn the predictive functions of different modalities can effectively address the modality laziness issue.

4.3.3 Analysis of Modality Gap

As demonstrated in Liang et al. [21], there exists a modality gap in multimodal learning, wherein different modality information is situated in two entirely separate regions within the embedding space. This modality gap exhibits a correlation with model performance, and increasing it can somehow enhance classification performance in multimodal learning [35]. To gain a deeper understanding of the performance improvements attributed to MLA, we visualize the modality gap between the text and vision modalities in the Food101 dataset in Figure 3. By comparing the performance with concatenation, MLA results in a larger modality gap, signifying that different modalities become more distinguishable and further lead to stronger performance. This further underscores the effectiveness of our proposed approach.
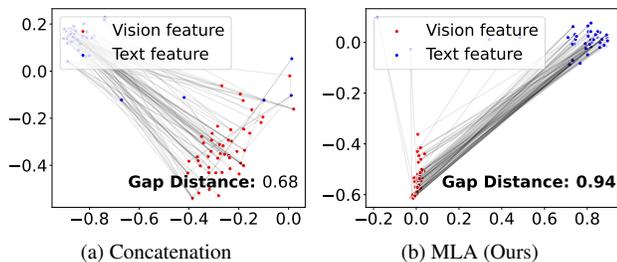

| (a) Concatenation | (b) MLA (Ours) |

Figure 3. Visualizations of the modality gap distance on the Food101 dataset.

4.3.4 Robustness Analysis of the Pre-trained Model

We evaluate the robustness of pre-trained models in this section. To do so, we utilize pre-trained features using Con-

trastive Language-Image Pretraining (CLIP) [28]. Specifically, we employ pre-trained CLIP encoders (ViT-B/32 pre-trained model loaded as both the image and text encoders) in the Food-101 dataset and apply MLA on this new encoders and present the results in Table 4. From Table 4, it is evident that MLA effectively mitigates modality laziness. This is demonstrated by a significant improvement in the performance of the subordinate visual modality, while simultaneously enhancing the fusion performance. These findings underscore the effectiveness of our method in multimodal representation learning.

Table 4. 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 (refer to Appendix A.4 for full results).

Method	Food-101		
	Image	Text	Multi
CLIP	63.07	83.98	93.07
CLIP + MLA (Ours)	**72.22**	**85.34**	**93.47**

5. Conclusion

In this paper, we propose MLA to address the challenge of imbalance optimization across different modalities in multimodal learning, which is referred to as "modality laziness." Specifically, our main idea is decomposing the traditional joint optimization strategy in multimodal learning into an alternating unimodal optimization strategy. To capture cross-modal information in the unimodal optimization process, we propose a gradient modification mechanism to avoid modality forgetting. Furthermore, during test time, MLA incorporates a dynamic fusion mechanism to enable effective multimodal fusion. MLA is also compatible with scenarios including learning with complete or missing modalities. The empirical results demonstrate the promise of MLA in mitigating modality laziness and further improving performance under five datasets with different types of modalities.

Acknowledgement

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. 2

[2] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1247–1255. JMLR.org, 2013. 6

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 609–617. IEEE Computer Society, 2017. 5

[4] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 153–160. MIT Press, 2006. 6

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMO-CAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008. 5

[6] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.*, 5(4):377–390, 2014. 5

[7] Shubai Chen, Xiaodong Yang, Yiqiang Chen, Hanchao Yu, and Hong Cai. Uncertainty-based fusion netwok for automatic skin lesion diagnosis. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1487–1492. IEEE, 2022. 2

[8] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024. 2

[9] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 8632–8656. PMLR, 2023. 1

[10] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. *arXiv preprint arXiv:2305.01233*, 2023. 2

[11] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multi-modal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 2

[12] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoen-coders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 5

[13] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 5

[14] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, Hawaii, USA, October 10-12, 2005*, pages 3437–3443. IEEE, 2005. 5, 6, 12

[15] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992. 3, 6

[16] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022. 2

[17] Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ish-tiaq. Improving multimodal accuracy through modality pre-training and attention. *CoRR*, abs/2011.06102, 2020. 1

[18] Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ish-tiaq. Improving multimodal accuracy through modality pre-training and attention. *arXiv preprint arXiv:2011.06102*, 2020. 2

[19] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomás Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5198–5204. AAAI Press, 2018. 5, 13

[20] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 2022. 1, 6

[21] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 2, 8

[22] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2302–2310, 2021. 3

[23] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. 3

[24] Netta Madvil, Yonatan Bitton, and Roy Schwartz. Read, look or listen? what's needed for solving a multimodal dataset. *CoRR*, abs/2307.04532, 2023. 2

[25] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. Sentiment analysis on multi-view social data. In *MultiMedia*

Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II, pages 15–27. Springer, 2016. 5

[26] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8228–8237. IEEE, 2022. 1, 2, 5, 13

[27] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press, 2018. 5, 13

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 8

[29] Dirk TM Slock and Thomas Kailath. Numerically stable fast transversal filters for recursive least-squares adaptive filtering. In *Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms*, pages 605–615. Springer, 1991. 4

[30] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Process. Lett.*, 28:1650–1654, 2021. 1, 2

[31] Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging. *arXiv preprint arXiv:2304.14933*, 2023. 1

[32] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. 1

[33] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4971–4980. IEEE Computer Society, 2017. 3, 6

[34] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023. 2

[35] Vishaal Udandarao. *Understanding and Fixing the Modality Gap in Vision-Language Models*. PhD thesis, Master's thesis, University of Cambridge, 2022. 8

[36] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015. 6

[37] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12692–12702. Computer Vision Foundation / IEEE, 2020. 1, 2

[38] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 2

[39] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multi-modal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015*, pages 1–6. IEEE Computer Society, 2015. 5

[40] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024. 2

[41] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the TVQA dataset. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. 2

[42] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 1

[43] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4400–4407. ACM, 2021. 6

[44] Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia*, pages 1–14, 2022. 6

[45] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022. 3

[46] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2402–2415, 2022. 3, 6

[47] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2402–2415, 2022. 6

[48] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. *arXiv preprint arXiv:2306.02050*, 2023. 1, 2, 5, 6, 13

[49] Xiaohui Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Le Xu, and Ruibo Fu. Adaptive fake audio detection with low-rank model squeezing. In *Proceedings of the Workshop on Deepfake Audio Detection and Analysis co-located with 32th International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, China, August 19, 2023*, pages 95–100. CEUR-WS.org, 2023. 4

[50] Xiaohui Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, and Chu Yuan Zhang. Do you remember? overcoming catastrophic forgetting for fake audio detection. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 41819–41831. PMLR, 2023.

[51] Xiaohui Zhang, Jiangyan Yi, Chenglong Wang, Chu Yuan Zhang, Siding Zeng, and Jianhua Tao. What to remember: Self-adaptive continual learning for audio deepfake detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19569–19577. AAAI Press, 2024. 4

[52] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 589–599. Springer, 2021. 3

[53] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2608–2618. Association for Computational Linguistics, 2021. 3, 6

[54] Wenhao Zheng, Dongsheng Peng, Hongxia Xu, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. Multimodal clinical trial outcome prediction with large language models. *arXiv preprint arXiv:2402.06512*, 2024. 2

[55] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[56] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 2

[57] Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. *CoRR*, abs/2210.15359, 2022. 6