

Supplementary Materials: ERL-MR

A Appendix for Modality Imbalance

THEOREM A.1. *For each modality m_i , during each iteration t , if m_i is the dominant modality, the performance measure $f(m_i)$ will converge more rapidly compared to the other modalities.*

PROOF. To facilitate our analysis, we begin by considering a multi-modal dataset \mathcal{D} comprising two distinct modalities, denoted as m_0 and m_1 . Thus, \mathcal{D} can be represented as $\mathcal{D} = \{(x_i^{m_0}, x_i^{m_1}, y_i)\}_{i=1}^N$, where N is the number of samples and $y = \{1, 2, \dots, C\}$ (C is the number of classes). Thus, we also have $f: x \rightarrow y$, i.e., the goal of f is to predict y from x .

Considering the general process of multimodal learning, we use encoders $\psi^0(\omega^{m_0}, \cdot)$ and $\psi^1(\omega^{m_1}, \cdot)$ to extract features from modality m_0 and modality m_1 , respectively, where ω^{m_0} and ω^{m_1} are the parameters of the encoder. After feature extraction, we need to perform a fusion operation on the extracted features. For simplicity, we use the vanilla fusion method, *concatenation*, to perform the fusion operation, which is widely used in multi-modal learning [2, 9]. Thus, the formal expression of the concatenation fusion operation is as follows:

$$f(x_i) = W \cdot [\psi^0(\omega^{m_0}, x_i^{m_0}); \psi^1(\omega^{m_1}, x_i^{m_1})] + b, \quad (1)$$

where $W \in \mathbb{R}^{C \times (d_{\psi^0} + d_{\psi^1})}$ is the weight matrix, $b \in \mathbb{R}^C$ denotes the parameters of the last linear classifier, and $f(x_i)$ is the logits output of the fusion model f . To facilitate the individual observation of the optimization process for each modality, the weight matrix W can be expressed as the combination of two distinct blocks: $[W^{m_0}, W^{m_1}]$. Eq. (1) can be rewritten as follows:

$$f(x_i) = W^{m_0} \cdot \psi^0(\omega^{m_0}, x_i^{m_0}) + W^{m_1} \cdot \psi^1(\omega^{m_1}, x_i^{m_1}) + b. \quad (2)$$

Subsequently, we define the loss function for the optimization process of each modality. Without any loss of generality, we illustrate this using the cross-entropy loss as an example, which is formally defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{f(x_i) y_i}}{\sum_{k=1}^C e^{f(x_i) k}}. \quad (3)$$

Generally, the stochastic gradient descent (SGD) algorithm is used to iteratively optimize Eq. (3). Then W^{m_0} (W^{m_1}) and the parameters encoder $\psi^0(\omega^{m_0}, \cdot)$ ($\psi^1(\omega^{m_1}, \cdot)$) are updated in the $t+1$ training round as follows:

$$\begin{aligned} W_{t+1}^{m_0} &= W_t^{m_0} - \eta \nabla_{W^{m_0}} \mathcal{L}_{CE} (W_t^{m_0}) \\ &= W_t^{m_0} - \eta \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)} \psi^0(\omega^{m_0}, x_i^{m_0}), \end{aligned} \quad (4)$$

$$\begin{aligned} \omega_{t+1}^{m_0} &= \omega_t^{m_0} - \eta \nabla_{\omega^{m_0}} \mathcal{L}_{CE} (\omega_t^{m_0}) \\ &= \omega_t^{m_0} - \eta \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)} \frac{\partial (W_t^{m_0} \cdot \psi_t^{m_0}(\omega_t^{m_0}, x_i^{m_0}))}{\partial \omega_t^{m_0}}, \end{aligned} \quad (5)$$

where η is the learning rate and t denotes the current training round. Inspired by previous work [2–4, 9] and combined with Eqs.

(3)–(4), we derive the following lessons: **(L1)** There is almost no correlation between the optimization of W and $\psi^0(\omega, \cdot)$ across different modalities; **(L2)** The term $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ related to the training loss can generalize the optimization correlation of W and $\psi^0(\omega, \cdot)$ between different modalities. The reason is that the encoder of each modality is optimized independently but the term $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ needs to be optimized for all modalities. To do this, we need to observe the gradient $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ for true label y_i and analyze it in conjunction with Eq. (2):

$$\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i) y_i} = \frac{e^{(W^{m_0} \cdot \psi^{m_0} + W^{m_1} \cdot \psi^{m_1} + b) y_i}}{\sum_{k=1}^C e^{(W^{m_0} \cdot \psi^{m_0} + W^{m_1} \cdot \psi^{m_1} + b) k}} - 1, \quad (6)$$

where $W^{m_0} \cdot \psi^{m_0} = W^{m_0} \cdot \psi^0(\omega^{m_0}, x_i^{m_0})$. Furthermore, to quantitatively evaluate the performance of each modality and the fusion model with respect to the training loss term, we provide the output logits for the ground truth label y as follows:

$$\begin{aligned} s^0 &= \text{softmax}(W^{m_0} \cdot \psi^0 + b/2)_y \\ s^1 &= \text{softmax}(W^{m_1} \cdot \psi^1 + b/2)_y \\ s^f &= \text{softmax}(W^{m_0} \cdot \psi^0 + W^{m_1} \cdot \psi^1 + b)_y. \end{aligned} \quad (7)$$

Discussion. (1) First, we discuss how each modality contributes to the $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ via $W \cdot \psi$. Referring to Eq. (6), it becomes evident that when a specific modality demonstrates superior performance, it will exert a predominant influence on the optimization of the training loss term $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ via $W \cdot \psi$. Consequently, in such a scenario, the modality with lower performance experiences a slower optimization pace due to its limited contribution to the training loss term, ultimately impeding the progress of the fusion model. It is worth noting that the above findings are consistent with the conclusions of previous work [2–4, 9]. **(2)** Second, we offer an insight that even if attempts are made to restrain dominant modalities to achieve modality balance, enhancing the overall performance of the fusion network proves to be a huge challenge. As indicated by Eq. (7), we understand that the performance of s^f is contingent on the combined performance of both modalities, rather than any single one in isolation. \square

B Appendix for Methodology

B.1 Explanations for ERL-MR

In order to verify the effectiveness of the proposed ERL-MR strategy in alleviating the modality imbalance phenomenon, we need to answer the following two questions:

- **Q1:** How does ERL-MR strategy improve the correlation and complementarity between different modality data?
- **Q2:** How does ERL-MR strategy efficiently constrain the optimization speed and direction of different modality data?

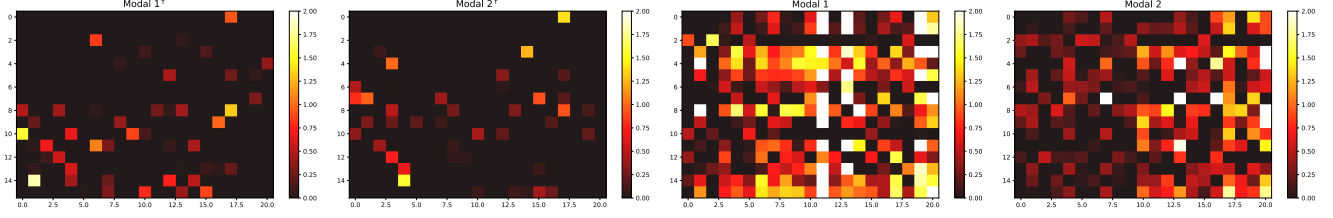


Figure 1: The feature representation maps of Modality 1 and Modality 2 obtained from the USC dataset at the 60th epoch. [†] indicates the ERL-MR strategy is applied.

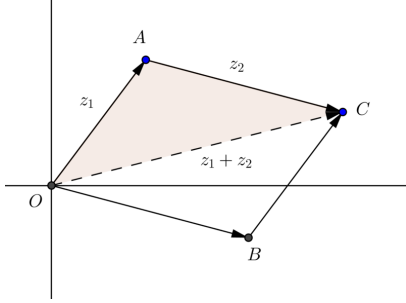


Figure 2: The visual understanding of the multi-modal interactions in ERL-MR.

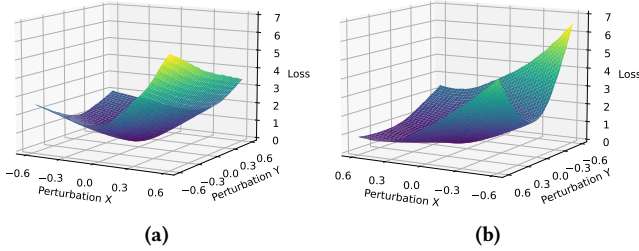


Figure 3: (a) Visualization of loss landscape when applying ERL-MR strategy. (b) Visualization of the loss landscape when the ERL-MR strategy is not applied.

Visualized Intuitive Explanation of Euler Feature Transformation. To address Q1, we introduce an Euler feature transformation design that effectively enhances the correlation and complementarity between different modalities. As illustrated in Fig. 1, Euler feature transformation not only substantially improves the feature correlation between different modalities but also reduces the redundancy of feature information. Specifically, we observe that the feature heatmaps of the two modalities generated by the ERL-MR strategy (*i.e.*, left of Fig. 1) are not only similar but also exhibit highly sparse. In contrast, the feature heatmaps of the two modalities that do not utilize our strategy (*i.e.*, right of Fig. 1) are quite different and possess highly redundant features. Next, we illustrate that Euler feature transformation can also enhance the complementarity between modality data. Let $z_1 = A_1 e^{j\theta_1}$ represent a complex feature vector from one modality, and $z_2 = A_2 e^{j\theta_2}$ represent a complex feature vector from another modality. According to the

triangle inequality for complex numbers, *i.e.*, $|z_1 + z_2| \leq |z_1| + |z_2|$, we have:

$$\begin{aligned} |A_1 e^{j\theta_1} + A_2 e^{j\theta_2}| &\leq |A_1 e^{j\theta_1}| + |A_2 e^{j\theta_2}| \\ &= |A_1| + |A_2|. \end{aligned} \quad (8)$$

The above inequality demonstrates that the modulus of the combined complex representation $|z_1 + z_2|$ is limited by the sum of the modulus amplitudes of the individual modalities ($|A_1| + |A_2|$). This inequality indicates that if A_1 and A_2 are significantly different, the combined modulus magnitude can not be larger than the sum of the individual modulus. Therefore, it shows that complex representation maintains the modulus complementarity across modalities. In addition, Fig. 2 also shows the geometric interpretation of the complementarity between modalities enhanced by Euler feature transformation.

Visualized Explanation of Multi-modal Constrained Loss. To address Q2, we visualize the loss landscape of MMCLoss and conventional loss in Fig. 3a, revealing that the convergence of MMCLoss is characterized by smooth and stable trajectories. In contrast, the convergence of the conventional loss not only exhibits significant jitter but also results in larger loss values, as shown in Fig. 3b. It implies that MMCLoss effectively constrains the gradient optimization direction between different modalities, leading to stable and smooth convergence. This success is attributed to MMCLoss’s ability to constrain the phases of the complex features.

B.2 ERL-MR Algorithm

Here, we briefly present the specific process of the ERL-MR strategy in Algorithm 1. At each epoch, a multi-modal training batch B_t is fed into the model as the input, and the features are extracted simultaneously for each modality m_i (Line 5). Lines 7-8 and Line 10 summarize the designed Euler feature transformation and feature compression, respectively. Line 11 illustrates the phase extraction of the complex features. Line 12 indicates the calculation of the cross-entropy loss at each training batch. Lines 14-16 outline the process of MMCLoss, while Line 17 details the model parameter update.

C Appendix for Evaluation

C.1 Dataset Information

USC Dataset [14]. The dataset comprises data collected from 14 users using 3-axis accelerometers and 3-axis gyroscopes. Two types of sensors have the same sampling rate of 100 Hz. In each 2-second

Algorithm 1: ERL-MR Algorithm

Input: The dataset $\mathcal{D} = \left\{ \left(x_i^{m_0}, x_i^{m_1}, \dots, x_i^{m_{M-1}}, y_i \right) \right\}_{i=1}^N$,
the initialized encoder parameters for each modality
 $\omega^0, \omega^1, \dots, \omega^{M-1}$, and the number of epoch E ;

Result: Optimal model parameters of each modality
 $\hat{\omega}^0, \hat{\omega}^1, \dots, \hat{\omega}^{M-1}$;

while training **do**
 for $t = 1, 2, \dots, E$ **do**
 Feeding-forward each mini-batch data $\mathcal{B}_t \in D$ to the model;
 for each modality m_i **in parallel** **do**
 Generate the embedding feature via the encoder
 $\psi^{m_i}(\omega^{m_i}, x^{m_i})$;
 // ---Complex Space Mapping---//;
 Obtain the real part feature $\psi_r^{m_i}(\omega^{m_i}, x^{m_i})$;
 Obtain the imaginary part feature
 $\psi_p^{m_i}(\omega^{m_i}, x^{m_i})$;
 // ---Feature Compression---//;
 Obtain $\psi_{\text{New}}^{m_i}(\omega^{m_i}, x^{m_i})$ via Eq. (8)–(10);
 Calculate the phase ϕ^{m_i} via Eq. (11);
 Calculate the $\mathcal{L}_{CE}^{m_i}$ via Eq. (13);
 end
 Calculate the $\sum_{m_i=0}^{M-1} \mathcal{L}_{\text{Cosine-similarity}}^{m_i}$ via Eq. (12);
 // ---MMCosine---//;
 Calculate the \mathcal{L}_{MMC} via Eq. (14);
 Update the model based on \mathcal{L}_{MMC} by using SGD;
 end
end

time window, a 600-dimensional vector is generated as each modality data. These vectors capture the data from the accelerometers and gyroscopes, providing a comprehensive representation of the respective modalities.

AVE Dataset [11]. The dataset is a specialized audio-visual video dataset for audiovisual event localization. It contains a total of 4,143 10-second videos collected from the YouTube website. The dataset covers a diverse set of 28 event classes and provides both auditory and visual trajectories, accompanied by secondary annotations. To support training and evaluation, the dataset is used in the experiment according to the predefined training and validation splits as defined in reference [11].

MHAD Dataset [7]. The dataset consists of data related to 11 different human behaviors, collected from a total of 12 subjects. Each frame of multi-modal data in the dataset includes two types of information: 3D accelerometer data and 35×3 dimensional skeletal points. In order to extract data samples, a sliding time window of 2 seconds is utilized for each subject. With this setup, approximately 330 samples are generated for each subject in the dataset.

Colored-gray MNIST Dataset (CGM) [5]. The dataset also known as CG-MNIST, is a synthetic dataset derived from the MNIST [6] dataset. CGM comprises pairs of images, where each pair consists of a gray-scale image and a monochrome image. The training set

of CGM contains 60,000 instances, where the monochrome images exhibit a strong color correlation with their respective numerical labels. Additionally, the testing set of CGM contains 10,000 instances with the weakened color correlation between the monochrome images and their labels compared to the training set.

FLASH Dataset [10]. The dataset consists of data collected by autonomous vehicles using GPS, lidar, and cameras. The data collection occurs at a frequency of 10 Hz. Each sample in the dataset includes a 64-dimensional RF ground truth and synchronized multi-modal sensor data. The dimensions of the sensor data are as follows: [1, 2] for GPS, [20, 20, 20] for lidar, and [3, 360, 640] for images. Its task is to select high-band sectors for mmWave beamforming in mobile V2X communication scenarios.

Alzheimer's Disease Monitoring (ADM) Dataset [8]. The dataset focuses on detecting Alzheimer's disease by analyzing 11 behavioral biomarkers in natural home environments. These biomarkers include activities such as cleaning living areas, taking medications, using mobile phones, writing, sitting, standing, getting in and out of chairs/beds, walking, sleeping, eating, and drinking. The three modal data of depth images, radar, and audio are obtained by sampling from the depth camera, mmWave radar, and microphone at sampling rates of 15 Hz, 20 Hz, 44 Hz, and 100 Hz, respectively.

CREMA-D Dataset [1]. The dataset comprises both facial and vocal emotional expressions, forming an audio-visual collection designed for emotion recognition research. It encompasses 7442 clips, representing emotional states categorized into six groups: happy, sad, anger, fear, disgust, and neutral. These clips are randomly split into 6698 samples for training and 744 samples for testing.

C.2 Baselines Information

PMR [2]. The PMR utilized prototypes to assess modality performance by measuring the sample distance between them. This enables continuous monitoring of modality imbalance during the training process. The introduced PCE loss improves prototype clustering, facilitates faster learning, reduces the dominance of uni-modality, alleviates the issue of modality imbalance, and solely relying on modality representation.

MMCosine [13]. The Multi-modal Cosine Loss (MMCosine) mitigates the norm dominance by trained uni-modal encoders through the modality-specific L_2 normalization on features and weights. It regulates the angle relationship between weights and features using cosine similarity, which alleviates alleviating optimization imbalance and facilitates facilitating multi-modal fine-grained learning.

OGM-GE [9]. The On-the-fly Gradient Modulation (OGM) strategy dynamically monitors the varying contributions of different modalities to the learning target during the training process. It adjusts the gradient accordingly to allocate more efforts to under-optimized modalities. Additionally, OGM introduces dynamically changing Gaussian noise to achieve Generalization Enhancement (GE), leading to significant performance improvements in under-optimized uni-modal representations.

Gradient-Blending [12]. The Gradient Blending (G-Blend) method minimizes the overfitting generalization ratio (OGR) by blending multiple supervision signals. To address the varied rates of overfitting and generalization across modalities, G-Blend employs joint training instead of a single optimization strategy. By calculating an

Table 1: Comparison of time overhead per epoch (seconds).

Method	Dataset			
	USC	MHAD	CGMNIST	AVE
None	2.2	3.3	24.4	112.2
G-Blend	2.5	3.8	24.5	112.9
OGM-GE	6.1	7.5	29.9	149.9
MMCosine	2.5	3.9	24.6	113.9
PMR	9.7	17.2	159.6	188.4
Ours	2.5	3.4	24.7	113.5

Table 2: Accuracy (%) comparison on ADM and Flash.

Method	ADM				Flash			
	Concat	Sum	FiLM	Gated	Concat	Sum	FiLM	Gated
G-Blend	38.2	35.5	36.3	36.3	57.4	56.5	57.2	56.9
OGM-GE	37.5	35.2	34.9	34.2	56.9	57.2	57.4	57.3
MMCosine	35.1	35.9	34.5	35.4	57.2	56.9	57.1	57.3
PMR	37.2	34.6	35.0	35.1	57.3	56.7	56.5	57.4
Ours	41.3	38.5	39.2	38.7	58.7	58.3	57.6	57.9

Table 3: Accuracy (%) comparison on CREAM-D.

Method	CREMA-D			
	Concat	Sum	Film	Gated
G-Blend	58.2	57.8	59.9	57.1
OGM-GE	58.9	58.4	60.9	60.0
MMCosine	57.4	58.8	59.1	61.3
PMR	58.4	57.3	61.2	59.3
Ours	69.0	66.8	67.2	66.3

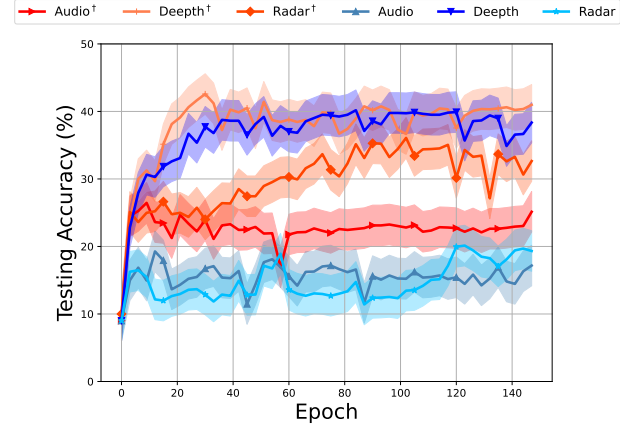
optimal blend among modalities, G-Blend effectively mitigates overfitting issues that arise from increased network capacity, resulting in improved performance.

C.3 Supplementary Experiments

In this part, we have conducted supplementary experiments to further validate the effectiveness of ERL-MR. These include comparative experiments on time overhead (Table 1), detailed comparative experiments on ADM and Flash datasets (Table 2), comparative experiments on the additional dataset CREAM-D (Table 3), and performance curves of modality balance on the ADM dataset (Fig. 4). The results indicate that ERL-MR exhibited better performance in both time overhead and performance comparison experiments. Furthermore, Fig. 4 illustrated the mitigating effect of ERL-MR on modality imbalance within the ADM dataset.

References

- [1] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

**Figure 4: The visual understanding of the multi-modal interactions in ERL-MR.**

- [2] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *Proc. of CVPR*.
- [3] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). In *Proc. of NeurIPS*.
- [4] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *Proc. of ICML*.
- [5] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proc. of CVPR*.
- [6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [7] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *Proc. of WACV*.
- [8] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Chen, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. In *Proc. of MobiSys*.
- [9] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proc. of CVPR*.
- [10] Batool Salehi, Jerry Gu, Debashri Roy, and Kaushik Chowdhury. 2022. FLASH: Federated learning for automated selection of high-band mmWave sectors. In *Proc. of INFOCOM*.
- [11] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proc. of ECCV*.
- [12] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proc. of CVPR*.
- [13] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. 2023. MMCosine: Multi-Modal Cosine Loss Towards Balanced Audio-Visual Fine-Grained Learning. In *Proc. of ICASSP*.
- [14] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proc. of UbiComp*.