COMPREHENSIVE MULTIMODAL KNOWLEDGE EXPLOITATION

YUNFENG FAN

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Computing

Comprehensive Multimodal Knowledge Exploitation

Yunfeng Fan

A thesis submitted in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**

April 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Yunfeng Fan

Abstract

Multimodal learning (MML) endeavors to simultaneously leverage the characteristics of various modalities to compensate their inherent limitations. In contrast to unimodal learning, MML can provide a clearer and more accurate perception of the target by removing redundancy and supplementing complementary information. Moreover, MML can enhance the robustness by reducing their reliance on a single modality. Models trained on multiple modalities are less susceptible to noise or errors in any single modality, resulting in more robust performance in real-world scenarios. However, the intrinsic heterogeneity between modalities makes it difficult to comprehensively utilize the multimodal information. Despite the great strides made yet, MML is still limited by three challenges that limit the exploitation of multimodal knowledge: 1) modality competition, 2) domain shift, and 3) distributed scenario. In this thesis, we explore effective ways to address the above challenges and design novel solutions to improve the learning efficiency of MML.

First, the joint training framework, which is commonly used in MML, inevitably falls into the notorious *modality competition*, making each modality under-explored. Specifically, modalities may interfere with each other, hindering the learning process especially for weak modalities. Therefore, in chapter 3, we introduce DI-MML, a novel *detached* MML framework designed to learn complementary information across modalities under the premise of avoiding modality competition. Specifically, DI-MML addresses competition by separately training each modality encoder with iso-

lated learning objectives. It further encourages cross-modal *interaction* via a shared classifier that defines a common feature space and employing a dimension-decoupled unidirectional contrastive (DUC) loss to facilitate modality-level knowledge transfer.

Second, we take the domain shift into consideration and study a more challenging task, multimodal domain generalization (MMDG) where models trained on multimodal source domains can generalize to unseen target distributions with the same modality set. Diverse modalities in real-world applications introduce more complex domain shifts, as the degree of domain shift varies across different modalities, significantly increasing the difficulty of addressing MMDG issue. Besides, previous domain generalization methods are specifically designed for unimodal setting and they are not compatible well in MMDG since the distinct properties between modalities leads to sub-optimal solutions. To bridge the gap, in chapter 4, we propose to construct consistent flat loss regions and enhance knowledge exploitation for each modality via cross-modal knowledge transfer. Innovatively, we turn to the optimization on representation-space loss landscapes instead of traditional parameter space, which allows us to build connections between modalities directly. Then, we introduce a novel method to flatten the high-loss region between minima from different modalities by interpolating mixed multi-modal representations.

Third, we consider a more complex MML scenario, multimodal federated learning (MFL), where multiple types of data is allocated on numerous distributed local devices. In this case, the diverse distribution heterogeneity of different modalities further increases the difficulty of exploiting multimodal knowledge effectively. However, existing federated learning (FL) frameworks employ client selection without taking into account the impact of modality differences across clients, as well as the modality bias. Thus, in chapter 5, we propose a novel Balanced Modality Selection framework for MFL (BMSFed) to overcome the bias. On the one hand, we incorporate a modal enhancement loss into local training to mitigate local imbalances by leveraging the aggregated global prototypes. On the other hand, we design a modality selec-

tion strategy to identify diverse subsets of local modalities, thereby ensuring global modality balance.

In summary, this thesis aims to maximize the extraction of knowledge from multiple modalities to achieve both efficiency and robustness across various complex scenarios. Extensive analysis and experimental evaluations show the performance advantages of our works with better performance over existing solutions.

Publications arising from the thesis

- [**NeurIPS**] **Yunfeng Fan**, Wenchao Xu, Haozhao Wang, Song Guo. "Cross-modal representation flattening for multi-modal domain generalization". The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

- [**MM**] **Yunfeng Fan**, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. "Detached and Interactive Multimodal Learning". The 32nd ACM International Conference on Multimedia, 2024.

- [**ECCV**] **Yunfeng Fan**, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. "Overcome Modal Bias in Multi-modal Federated Learning via Balanced Modality Selection". The 18th European Conference on Computer Vision, 2024.

- [**CVPR**] **Yunfeng Fan**, Wenchao Xu, Haozhao Wang, Junxiao Wang, Song Guo. "PMR: Prototypical modal rebalance for multimodal learning". IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

- [**INFOCOM**] Haozhao Wang, Wenchao Xu, **Yunfeng Fan**, Ruixuan Li, Pan Zhou. "AOCC-FL: Federated Learning with Aligned Overlapping via Calibrated Compensation". IEEE INFOCOM 2023-IEEE Conference on Computer Communications, 2023.

- [**AAAI**] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, **Yunfeng**

Fan. "Non-exemplar Online Class-Incremental Continual Learning via Dual-Prototype Self-Augment and Refinement". 38th AAAI Conference on Artificial Intelligence, 2024.

Acknowledgments

Completing this Ph.D. dissertation has been a challenging yet deeply rewarding journey, one that would not have been possible without the support, guidance, and encouragement of many remarkable individuals. I would like to take this opportunity to express my heartfelt gratitude to everyone who has contributed to this achievement.

First and foremost, I am deeply grateful to my parents, whose unconditional love and unwavering support have been the foundation of my academic journey. Their faith in me, their sacrifices, and their encouragement during both the highs and lows have been my greatest source of strength throughout this process. I owe this milestone to their endless care and belief in my potential.

I would also like to express my sincere gratitude to my supervisors, Prof. Wenchao Xu, Prof. Song Guo and Prof. Xiao Bin. Their insightful guidance, patience, and dedication have not only helped me navigate the complexities of my research but also broadened my intellectual horizons. Their high standards and constructive feedback have challenged me to grow as a scholar in my work. In particular, I am profoundly grateful to Prof. Wenchao Xu, whose mentorship has been a cornerstone of my academic journey. His deep expertise, visionary guidance, and unwavering support have not only shaped the direction of my research but also inspired me to think critically and approach challenges with confidence. I am also deeply thankful to Prof. Song Guo for his invaluable insights and rigorous approach to research, which have significantly enriched my academic development. His constructive advice and high expectations have pushed me to refine my work and achieve a higher level of excellence.

Finally, I want to thank my friends and colleagues. Their camaraderie, support, and collaboration have made the long hours of research and study more bearable and even enjoyable. I am truly grateful for the memories we have created and the encouragement we have given each other along the way.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In summary, this thesis explores effective ways to comprehensively exploit the knowledge from multimodal data under various formidable challenges. This is an important research problem with a wide range of applications in machine learning. In this chapter, we first give a brief overview of the research problems in §1.1. Then we highlight the main contributions of this thesis in §1.2. Finally, §1.3 outlines the thesis organization.

1.1 Overview

Artificial Intelligence (AI) will be one of the techniques that benefit from the vast amount of multimodal data expected in the coming years. However, data from different modalities are typically acquired using distinct sensors, resulting in disparate forms of data representation. For instance, visually perceived content is often represented as images or videos, whereas sound is typically expressed as spectra, and language is commonly represented as text. These diverse representations exacerbate the heterogeneity between modalities, posing significant challenges for machine understanding. In this context, multimodal learning (MML) [80, 95, 92] has emerged as a promising approach to enable machines to perceive and comprehend data from various modalities. In MML, we need to train a neural network to process data from various modalities, which could exhibit strong heterogeneity between them, e.g., images are commonly represented as pixels, while audio is generally spectral data, and text is even a human-created type of information that does not exist in nature. The vastly different manifestations of these modalities make it challenging to integrate their information to complete tasks.

To ensure the full knowledge exploitation for different modalities, two key components in MML should be well-considered: *representation learning* and *fusion*. Specifically, representation learning [41, 70] refers to extracting meaningful and informative representations from multiple modalities of data, which are expected to capture the underlying semantic meaning and relationships between modalities. A good representation learning strategy enables the model to generalize well across modalities and perform effectively on various tasks, even when the data is severe complex and heterogeneous. Fusion [37, 38] aims to combine information from multiple modalities to perform a prediction for specific tasks. Fusion techniques tend to leverage the complementary nature of different modalities to improve overall performance or understanding. The joint utilization of representation learning and fusion can effectively explore and exploit the complementary information of the different modalities. Therefore, they play crucial roles in enabling machines to understand and interpret multimodal information in MML.

Although multimodal data contains rich and diverse information, it is still challenging to effectively and fully extracting this knowledge. For instance, current multimodal learning approaches tend to focus more on complex training paradigms or model architectures [2, 119, 47], while paying insufficient attention to the interactions between modalities during the learning process. As discussed in [85], different modalities exhibit substantial variability during training, manifesting as performance imbalances. This discrepancy arises from various factors, including the quantity of data, the compatibility between the modalities themselves, and the suitability of the model architecture. Fundamentally, the competition between modalities during multimodal training lies at the core of this issue, where the gradients of one modality interfere with others, ultimately preventing each modality from being adequately learned. Second, the emergence of requirements to complete multimodal tasks in different environments highlight the need to address multimodal domain generalization (MMDG), but the diverse degrees of shifts and distinct model architectures for different modalities make it extremely challenging to simultaneously improve the generalization performance of all modalities. Beyond those, the distribution heterogeneity caused by data being allocated across various devices is another aspect of MML that needs to be taken into consideration. The aforementioned issues highlight the difficulty of fully leveraging multimodal knowledge, motivating us to design efficient and robust multimodal training strategies.

Therefore, in this thesis, we aim to maximize the extraction of knowledge from multiple modalities to achieve both efficiency and robustness across various complex scenarios. In the first part, we introduces DI-MML, a novel detached MML framework designed to learn complementary information across modalities under the premise of avoiding modality competition. In the second part, we construct shared representation space instead of traditional parameter space to build connections between modalities directly, and propose to flatten their high-loss representation regions by interpolating mixed multi-modal representations. In the third part, we focus on the multimodal federated learning framework (MFL) and propose a novel *balanced modality selection* scheme instead of client selection to comprehensively exploit all modalities.

The rest of this report presents the research motivation, method designs, and evaluation of my existing works, as well as the future research schedule to complete my thesis and research programme. To make it easier to understand, we give a table of acronyms in Table 1.1.

Table 1.1: Acronyms

Acronym	Full Form
AI	Artificial Intelligence
BMSFed	Balanced Modality Selection for Multimodal Federated Learning
CL	Contrastive Learning
CMRF	Cross-Modal Representation Flattening
DG	Domain Generalization
DI-MML	Detached and Interactive Multimodal Learning
DNN	Deep Neural Network
DUC	Dimension-decoupled Unidirectional Contrastive
ERM	Empirical Risk Minimization
FL	Federated Learning
LLM	Large Language Model
MCRL	Multimodal Contrastive Representation Learning
ME	Modal Enhancement
MFL	Multimodal Federated Learning
MLLM	Multimodal Large Language Model
MMDG	Multimodal Domain Generalization
MML	Multimodal Learning
RAG	Retrieval-Augmented Generation
SAM	Sharpness-Aware Minimization
SMA	Simple Moving Average
UML	Unimodal Learning

1.2 Thesis Contribution

We briefly summarize our contributions below.

1. **Avoid Modality Competition via Detached Multimodal Learning**

   To tackle modality competition, we introduce DI-MML, a novel *detached* MML framework designed to learn complementary information across modalities under the premise of avoiding modality competition. DI-MML separately trains each modality encoder with isolated learning objectives and further encourages cross-modal *interaction* via a shared classifier and specifically designed dimension-decoupled unidirectional contrastive (DUC) loss. Extensive experiments conducted on various datasets show the superior performance of our proposed method.

2. **Achieving Consistent Flat Loss Regions on Representation Space**

   We are the first to extend the unimodal flatness analysis to MMDG. We then construct shared representation space instead of parameter space to build connections between modalities directly and propose to flatten high-loss representation regions between modalities by interpolating mixed multi-modal representations and performing knowledge distillation to regularize the learning of each modality. Extensive experiments verify the effectiveness and superiority of our framework on two benchmark datasets of EPIC-Kitchens and HAC.

3. **Balanced Modality Selection on MFL**

   We reveal that uni-modal training on some clients may contribute more to the global model than multi-modal training. Based on the analysis, we propose a novel Balanced Modality Selection scheme for MFL (BMSFed) to comprehensively exploit all modalities via a modal enhancement loss and representative modality selection to overcome the global modal bias. We conduct comprehensive experiments on various datasets, and considering the statistical heterogeneity and modality incongruity problems in MFL, to validate the superiority of our BMSFed.

1.3 Thesis Outline

The rest of this thesis consists of five chapters and organized as follows. In §2, we review the background knowledge on MML. In §3, we present DI-MML, a novel detached MML framework designed to learn complementary information across modalities. In §4, we propose to construct consistent flat loss regions and enhance knowledge exploitation for each modality via cross-modal knowledge transfer. In §5, we propose a novel Balanced Modality Selection framework for MFL (BMSFed) to overcome the modal bias. We summarize this thesis and provides some potential future research directions in §6.

Chapter 2

Background Review

2.1 Multimodal Learning

Multimodal learning (MML) [88] has emerged as a promising approach to enable machines to perceive and comprehend data from various modalities. The field of MML has undergone significant evolution, progressing from end-to-end training and pre-training to fine-tuning on current foundation models. These models have demonstrated an improved ability to perceive and understand the world, even surpassing human performance in certain scenarios.

Unimodal learning (UML) focuses on exploring and exploiting knowledge from a single modality. Despite the significant successes of UML, effectively processing the diverse modal information that surrounds us is crucial, and leveraging multimodal information is essential for various applications and real-world scenarios. In contrast to UML, multimodal learning (MML) is capable of handling different types of data, offering several advantages. Firstly, data often naturally exists in a multimodal form. The representation of an object encompasses not only visual information, but also other types of information, including smell or touch. By incorporating multiple modalities, MML can provide a more comprehensive and accurate characterization of the target. Furthermore, multiple modalities contain richer information, comprising both common and specific features. By removing redundancy and supplementing complementary information, MML can provide a clearer and more accurate perception of the target. Secondly, MML can enhance the robustness by reducing their reliance on a single modality. Models trained on multiple modalities are less susceptible to noise or errors in any single modality, resulting in more robust performance in real-world scenarios. Thirdly, MML facilitates better generalization to unseen data or tasks by leveraging complementary information from multiple modalities. Models trained on diverse modalities can capture underlying patterns that may not be apparent in any single modality, leading to improved generalization performance. Additionally, in domains where data in a single modality is sparse or insufficient for training accurate models, MML can leverage data from multiple modalities to improve performance. This is particularly beneficial in scenarios where collecting data in one modality may be more challenging or expensive.

2.2 Imbalanced Multimodal Learning

Several recent studies [106, 24, 111] have shown that many multimodal deep neural networks (DNNs) cannot achieve better performance compared to the best single-modal DNNs. This phenomenon is termed as “modality competition” (we also use modality imbalance to describe it consequence in this thesis). As defined in [49], it means that during joint training, multiple modalities will compete with each other. **Only a subset of modalities which correlate more with their encoding networks random initialization will win and be learned with other modalities failing to be explored**. Wang et al. [106] found that different modalities overfit and generalize at different rates and thus obtain suboptimal solutions when jointly training them using a unified optimization strategy. Peng et al. [85] proposed that the better-performing modality will dominate the gradient update while suppressing the learning process of the other modality. Furthermore, it has been reported that multimodal DNNs can exploit modal bias in the data, which is inconsistent with the expectation of exploiting cross-modal interactions in VQA [52, 40, 111].

The core of current methods for addressing modality competition in joint training involves mitigating the imbalance between modalities by modulating the learning progress of different modalities. These methods can be categorized into three types based on what they operate on: modality-specific learning rate adjustment, gradient modulation, and feature-level optimization.

1) *Modality-specific learning rate adjustment:* The primary manifestation of modality competition is the inhibition in some modalities, leading to slow learning progress. In deep learning, the learning rate directly adjusts the network's learning pace. Thus, the central idea of modality-specific learning rate adjustment is to balance learning progress by adjusting the learning rate of each modality—accelerating slower-learning modalities or decelerating faster-learning ones. Yao et al. [126] revealed that optimal learning rates vary across modalities and assigned distinct learning rates to different modalities in late-fusion models. Similarly, [97, 35] optimized learning rates in comparable ways. Additionally, learning rates can be modulated by controlling the modalities involved in training. For example, [114] proposed DropPathway, which randomly drops the audio pathway during training iterations with a certain probability, thereby slowing the audio modal learning and making its dynamics more compatible with the visual pathway.

2) *Gradient modulation:* Gradients are fundamental to model updating, so direct control of gradients allows for finer tuning of learning. Gradient modulation balances modal behavior by imposing constraints on the divergences of gradients. [106] was the first to address modality competition with gradient modulation by dynamically adjusting unimodal and multimodal gradients. Subsequent works, such as [85] and [30], approached this by modulating gradient magnitudes and directions, respectively. [110] further introduced a sample-level modality valuation metric to evaluate

the sample-wise contribution of each modality and apply finer grained regulation. This allows them to observe that the modality discrepancy can vary across different samples, beyond just the global contribution discrepancy at the dataset level. While these methods aim to balance performance between modalities, [64] argued that consistent modality-wise performance is suboptimal due to inherent heterogeneity in ability between modalities, and proposed instead to adjust gradients to optimize each modal capacity as in unimodal training.

3) *Feature optimization:* Features represent the knowledge learned by networks to complete specific tasks, and modality competition affects feature quality. For instance, [86] found that well-behaved networks tend to output features with large L2 norms, prompting a method to reduce norm gaps between modalities. Additionally, varying data quality across modalities can exacerbate competition. To address overreliance on noisy audio data in audio-visual speech recognition tasks, Chen et al. [12] leveraged visual modality-specific representations to complement audio inputs. Similarly, [25] proposed using better features to facilitate feature learning for a particular modality. Unlike [12], which learns from a better modality, [25] learns from a superior model of the same modality. Considering that neural networks tend to learn easy patterns during early training stage and gradually learn more complex features, [137] proposed to adjust the difficulty of participating multimodal data during learning process, so the feature learning is also arranged from simple to hard. Fan et al. [29] further ensured the beneficial interactions between modalities through detached knowledge transfer.

Although a certain degree of improvement is achieved, such approaches do not impose the intrinsic motivation of improvement on the slow-learning modality, making the improvement of this modality a passive rather than an active behavior. Besides, the interference from other modalities will hinder the improvement by modulation based on the fused modality data. Furthermore, the application scenarios of these methods are limited by fusion methods or model structures.

2.2.1 Multimodal Federated Learning

Federated learning (FL) [75] works to jointly train a global model with a large number of clients while preserving privacy. To tackle the statistical heterogeneity in FL, FedProx [66] adds a proximal term to the objective that helps to improve the stability. FedProto [98] shares the abstract class prototypes instead of the gradients between server and clients to regularize the training of local models. FedNH [20] improves the generalization of the global model by distributing class prototypes uniformly in the latent space to solve the class imbalance setting. However, current methods mainly focus on unimodal settings, which makes it hard to satisfy the increasing demand for multimodal scenarios.

Only limited attempts have been made to solve multimodal tasks in FL (MFL). FedIoT [134] is a multimodal FedAvg [75] algorithm to extract correlated representations from local autoencoders. FedMSplit [14] focuses on modality heterogeneity in MFL. It splits local models into several components and aggregates them by the correlations amongst multimodal clients according to a dynamic and multi-view graph structure. CreamFL [128] comprehensively takes into account statistical heterogeneity, model heterogeneity and task heterogeneity in MFL, and uses knowledge distillation [45] with contrastive learning [15] via a public dataset. Although the literature considers various cases in MFL, the modality imbalance, which is vital in multimodal learning, has been ignored. In this thesis, we mainly try to tackle the challenges of combining modality imbalance and input heterogeneity in MFL.

Chapter 3

Detached and Interactive Multimodal Learning

Recently, Multimodal Learning (MML) has gained significant interest as it compensates for single-modality limitations through comprehensive complementary information within multimodal data. However, traditional MML methods generally use the *joint* learning framework with a uniform learning objective that can lead to the modality competition issue, where feedback predominantly comes from certain modalities, limiting the full potential of others. In response to this challenge, this section introduces DI-MML, a novel *detached* and *interactive* MML framework designed to learn complementary information across modalities under the premise of avoiding modality competition. Specifically, DI-MML addresses competition by separately training each modality encoder with isolated learning objectives. It further encourages cross-modal *interaction* via a shared classifier that defines a common feature space and employing a dimension-decoupled unidirectional contrastive (DUC) loss to facilitate modality-level knowledge transfer. Additionally, to account for varying reliability in sample pairs, we devise a certainty-aware logit weighting strategy to effectively leverage complementary information at the instance level during inference. Extensive experiments conducted on audio-visual, flow-image, and front-rear view datasets show the superior performance of our proposed method.

3.1 Introduction

Multimodal learning (MML) has emerged to enable machines to better perceive and understand the world with various types of data, which has already been applied to autonomous driving [115], sentiment analysis [56], anomaly detection [107], etc. Data from different modalities may contain distinctive and complementary knowledge, which allows MML outperforms unimodal learning [48]. Despite the advances in MML, fully exploiting the information from multimodal data still remains challenging.

Recent studies [106, 49] have found that the unimodal encoder in MML underperforms its best unimodal counterpart trained independently. Huang et al. [49] attribute the cause of this phenomenon to *modality competition*, where the dominant modality hinders the learning of other weak modalities, resulting in imbalanced modality-wise performance. Existing solutions [85, 35, 126] mainly try to modulate and balance the learning paces of different modalities, which generally follow the joint training framework and a uniform learning objective is employed for all modalities, as shown in Figure 3.1. However, according to [30], the fused uniform learning objective is actually the reason for modality competition since the backward gradient predominantly comes from certain better modalities, hindering the learning of others, as illustrated in Figure 3.2. Meanwhile, [25] has declared that despite the competition between modalities, the interactions in joint training can facilitate the exploitation of multimodal knowledge. Therefore, existing solutions are caught in the dilemma of mitigating competition and facilitating interactions, where the competition issue has not been eradicated, limiting further improvements in multimodal performance.

In this section, we empirically reveal that eliminating modality competition may be

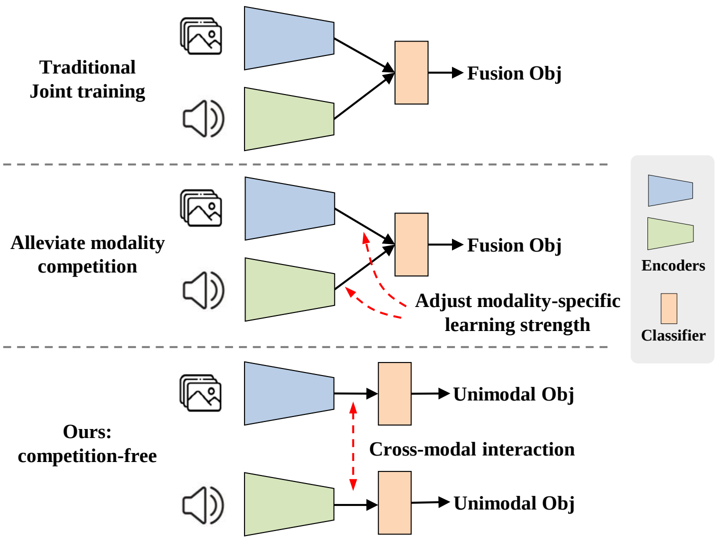


Figure 3.1: The difference between previous methods with ours. Only our method abandons the uniform fusion objective and updates each modal network with isolated objectives.

more critical for multimodal learning, which motivates us to design a competition-free training scheme for MML. Therefore, we decide to abandon the joint training framework and construct a novel *detached* learning process via assigning each modality with isolated learning objectives. Although the naive detached framework, i.e., performing unimodal training independently, could avoid modality competition, it still suffers from the following two challenges, limiting its further improvement.

- **Disparate feature spaces**. The intrinsic heterogeneity between modalities usually requires different processing strategies as well as model structures, which may lead to disparate feature spaces based on independent unimodal training and then pose a great challenge on fusing the extracted multimodal knowledge.

- **Lack of cross-modal interactions.** The cross-modal interactions can help to facilitate the exploitation of multimodal knowledge. However, independent unimodal training insulates the interactions for both encoder training and multimodal prediction process, limiting the learning and exploitation of multimodal complementary information.

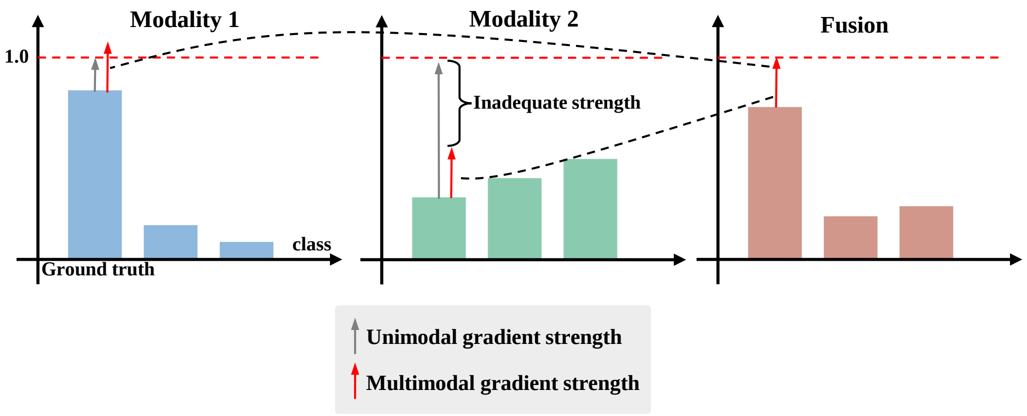


Figure 3.2: Modality competition comes from uniform learning objective. The columns represent predicted probabilities for each class. The fused prediction is dominated by modality 1 (better), resulting in a significant gap between the fusion gradient and the gradient needed for modality 2 (weak).

To address all above issues, we propose a novel detached and interactive multimodal learning (DI-MML) that achieves cross-modal **I**nteractions under the **D**etached training scheme. Unlike independent unimodal training, we first apply an additional shared classifier to regulate a shared feature space for various modalities, alleviating the difficulty on fusion process. To encourage cross-modal interactions during encoder training, we propose a Dimension-decoupled Unidirectional Contrastive (DUC) loss to transfer the modality-level complementary knowledge. We introduce the dimension-wise prediction to evaluate the discriminative knowledge for each dimension and then divide feature dimensions into effective and ineffective groups, enabling the complementary knowledge transfer within modalities and maintaining the full learning of each modality itself. Further, to enhance interactions during multimodal prediction, we then freeze the learned encoders and train a fusion module. Considering that there may be reliability disparities between modalities in sample pairs, we devise a certainty-aware logit weighting strategy during inference so that we can fully utilize the complementarities at the instance level.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to completely avoid modality competition while ensuring complementary cross-modal interactions in MML. We propose a novel DI-MML framework that trains each modality with isolated learning objectives.

- We design a shared classifier to regulate a shared feature space and a Dimension-decoupled Unidirectional Contrastive (DUC) loss to enable sufficient cross-modal interactions, which exploits modality-level complementarities.

- During inference, we utilize the instance-level complementarities via a certainty-aware logit weighting strategy.

- We perform extensive experiments on four datasets with different modality combinations to validate superiority of DI-MML and its effectiveness on competition elimination.

3.2 Related Work

3.2.1 Modality Competition in MML

Multimodal learning is expected to outperform the unimodal learning scheme since multiple signals generally bring more information [48]. However, recent research [106] has observed that the multimodal joint training network underperforms the best unimodal counterpart. Besides, even if the multimodal network surpasses the performance of the unimodal network, the unimodal encoders from multimodal joint training perform worse than those from unimodal training [24, 111, 112]. This phenomenon is termed as “modality competition” [49], which suggests that each modality cannot be fully learned especially for weak modalities since there exists inhibition between them. Researchers have proposed various methods to address this challenge, including gradient modulation [85, 30], learning rate adjustment [126, 97], knowledge

Table 3.1: The modality competition analysis on CREMA-D, AVE and UCF101. The metric is the top-1 accuracy (%). 'Audio', 'Visual', 'Flow' and 'Image' denote the corresponding uni-modal performance in each dataset. 'Multi' is the multimodal performance. 'Uni1' and 'Uni2' mean unimodal training based on audio and visual data respectively for CREMA-D and AVE, while flow and image respectively for UCF101.

Dataset	CREMA-D [9]			AVE [99]			UCF101 [94]		
Method	Audio	Visual	Multi	Audio	Visual	Multi	Flow	Image	Multi
Uni1	65.59	-	-	**66.42**	-	-	55.09	-	-
Uni2	-	78.49	-	-	46.02	-	-	42.96	-
Joint training	61.96	38.58	70.83	63.93	24.63	69.65	33.78	37.54	51.92
MM Clf	65.59	78.49	78.09	66.42	46.02	72.39	55.09	42.96	60.67
Preds Avg	65.59	78.49	82.66	66.42	46.02	69.40	55.09	42.96	64.43
CM Dist	63.17	77.28	82.93	62.94	41.79	67.41	54.30	42.93	64.45
Ours	**66.67**	**78.90**	**83.74**	64.18	**49.25**	**75.37**	**58.52**	**48.59**	**65.79**

distillation [25], etc. Despite their improvement, the competition phenomenon still exists since they insist on leveraging joint training scheme with a uniform learning objective, which is the culprit for modality competition [30]. The preserved competition greatly limits the improvement of multimodal performance. In this section, we aims to design a competition-free MML scheme which assigns isolated learning objectives to each modality without mutual inhibition, and guarantee the cross-modal interaction simultaneously.

3.2.2 Contrastive Learning in MML

Contrastive learning (CL) [15] aims to learn an embedding space where positive samples are clustered together while negative samples are pushed apart. Traditionally, CL has been applied to unimodal scenarios, e.g., self-supervised learning [54, 43], domain generalization [125, 60] and few-shot learning [72, 123]. In recent years, multimodal contrastive representation learning (MCRL) [87, 65] has been proposed to learn a shared feature space where the semantically aligned cross-modal representations are acquired. In MCRL, the paired multimodal samples are viewed as positive samples while the mismatched sample pairs are considered as negative samples. The cross-modal contrastive loss aims to pull the positive representations close in the instance level. MCRL has achieved great success yet. Multimodal pretrained models [36] emerged based on it, e.g., the vision-language models UniCL [122], FILIP [124], audio-text model CLAP [27] and audio-visual model CAV-MAE [39]. However, these methods are designed to align shared information in different modalities while overlooking the learning about the modality-specific and complementary features. In this section, we aim to achieve cross-modal interaction during the unimodal learning process via the complementary knowledge transfer based on CL.

3.3 Modality Competition Analysis

Let x be a data sample and $y = [K]$ be the corresponding label. Without loss of generality, we consider two input modalities $x = [x^1, x^2]$. In MML, we generally use two encoders ϕ^1, ϕ^2 to extract features of each modality: $\boldsymbol{h}^1 = \phi^1\left(\theta^1, \boldsymbol{x}^1\right)$ and $\boldsymbol{h}^2 = \phi^2\left(\theta^2, \boldsymbol{x}^2\right)$, where θ^1 and θ^2 are the parameters of encoders. And then, a fusion module is employed to integrate the information from two modalities and make predictions, i.e. $\psi\left(\boldsymbol{h}^1, \boldsymbol{h}^2\right)$, where ψ denotes the fusion and prediction function. The overall function of multimodal model can be written as $f\left(x\right) = \psi\left(\phi^1\left(x^1\right), \phi^2\left(x^2\right)\right)$. Therefore, the cross-entropy loss for multimodal classification is:

$$\mathcal{L}_{CE}(x) = -\log \frac{\exp\left(f\left(x\right)_y\right)}{\sum_{k=1}^{K} \exp\left(f\left(x\right)_k\right)} \tag{3.1}$$

This is a uniform learning objective for both modalities. MML is expected to exploit the complementary information of all modalities to outperform unimodal learning, but the modality competition phenomenon limits the performance improvement of MML since the dominant modality will inhibit the learning process of other modalities. As demonstrated in Table 3.1, the unimodal performance from the traditional

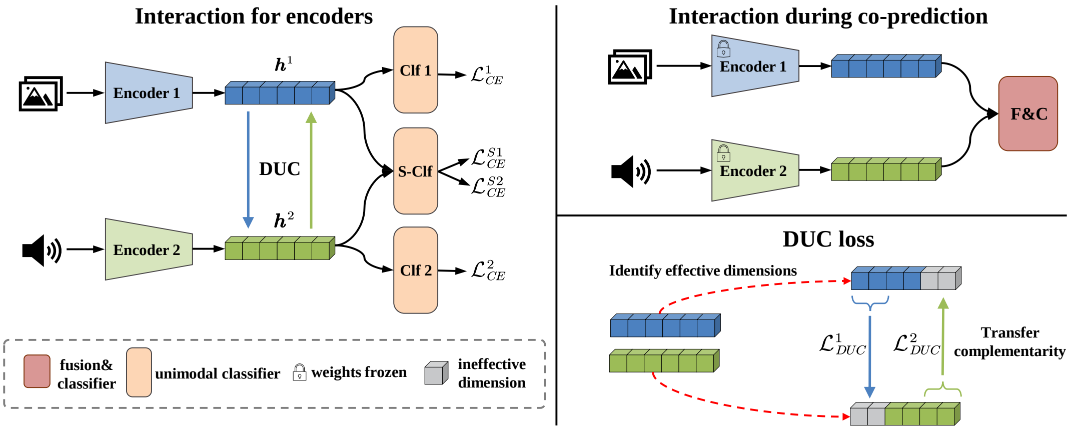


Figure 3.3: Overall framework of DI-MML. The encoders of each modality are trained with isolated learning objectives. The connections and interactions between modalities during encoder training are enabled by shared classifier and DUC loss.

multimodal joint training severely underperforms the results from corresponding unimodal training.

Although several methods [106, 85, 97] have been proposed to alleviate the modality competition, we find that the culprit behind, a uniform learning objective for both modalities, has not been resolved. According to the loss function Eq. 3.1, we can obtain the gradient of the softmax logits output with ground-truth label y:

$$\frac{\partial \mathcal{L}_{CE}}{\partial f\left(x\right)_y} = \frac{\exp\left(f\left(x\right)_y\right)}{\sum_{k=1}^{K} \exp\left(f\left(x\right)_k\right)} - 1 \tag{3.2}$$

which is the gap between the predictive probability on ground truth with the value 1. If one modality performs better (i.e., the needed gradient strength should be low) and dominates the fusion feature, the strength of generated gradient with the uniform learning objective could be weak, which cannot satisfy the requirement of greater gradient strength for the weak modality, as illustrated in Figure 3.2. Therefore, *removing the uniform learning objective for encoder training* is the key to eliminating modality competition.

Intuitively, we can perform the detached unimodal learning for each encoder independently and then fuse their outputs (features or logits). As shown in Table 3.1, we fix the pretrained unimodal learned networks and fuse their information in two ways: (1) MM Clf, train a multimodal linear classifier with the output features; (2) Preds Avg, average the prediction of each modality. It is clear that they can achieve impressive improvement compared with joint training despite the restricted cross-modal interactions, indicating the necessity to eliminate competition in MML. However, there still remain some challenges. Firstly, due to the heterogeneity between modalities, independent unimodal training may lead to disparate latent feature spaces. The correlations between modalities are ignored, making it difficult to fuse information effectively. For example, MM Clf on CREMA-D and UCF101 is worse than Preds Avg since the heterogeneous feature spaces hinder the feature fusion. Secondly, according to [25], the cross-modal interactions in joint training can help to explore the complementary information that is hard to be learned with unimodal training. Independent encoder training blocks cross-modal interactions, thus, limiting the use of multimodal complementary knowledge. Here we apply naive cross-modal logit distillation in independently unimodal training, namely CM Dist, to achieve inter-modal knowledge transfer, enabling the multimodal interactions via prediction with multimodal data as in joint training. It can be seen that CM Dist is better than MM Clf and Preds Avg on CREMA-D and UCF101, showing the potential of cross-modal knowledge transfer for multimodal interactions. Nonetheless, the naive distillation does not consider the heterogeneity between the modalities so it does not work well always (perform worse on AVE), which motivates us to design more delicate cross-modal interactive behavior.

We then present our method in next subsection, which not only solves all of the above challenges but achieves consistent improvement for various datasets on both multi- and uni-modal accuracy.

3.4 Methodology

3.4.1 Detached and Interactive MML

According to the above discussion, we separately train each modality's encoder to avoid modality competition. Meanwhile, we enable cross-modal interactions during the encoder training and fusion process, as well as inference, to exploit the complementary information between different modalities. The details are given below and the overall framework is shown in Figure 3.3.

Detached unimodal training. The network of each modality is updated only according to its own data and learning objectives, and there is no fusion during the update of encoders. Encoders ϕ^1, ϕ^2 are equipped with corresponding classifiers ψ^1 and ψ^2. Therefore, the logit output of modality i is $\boldsymbol{z}^i = f^i\left(x^i\right) = \psi^i\left(\phi^i\left(x^i\right)\right)$, $i \in \{1, 2\}$. The classification loss $\mathcal{L}_{CE}^i\left(x^i\right)$ of each modality is independent with each other, exploiting informative knowledge for classification.

Interaction during encoder training. To address the disparate feature spaces, we use a shared linear classifier (S-Clf) for different modalities to regulate the consistent feature space. Given the extracted features $\boldsymbol{h}^i$, the logit output through the shared classifier is $\boldsymbol{sz}^i = W\boldsymbol{h}^i + b$, where $W = [W_1, \cdots, W_K] \in \boldsymbol{R}^{d\times K}$, $b \in \boldsymbol{R}^d$ are the parameters of S-Clf and d is the feature dimension. According to [93, 71], the paired features $\boldsymbol{h}^1, \boldsymbol{h}^2$ with label y are optimized to maximize the similarity between them with the y-th vector W_y, and hence, S-Clf forces two modalities to locate at the same feature space using W_y as the anchor. The corresponding loss for each modality is denoted as $\mathcal{L}_{CE}^{Si}\left(x^i\right)$.

Then, we need to enable the cross-modal interaction to exploit the complementary information. According to the analysis in Section 3.3, cross-modal knowledge transfer is a promising way for interactions. Considering the gap between modalities [120], we intend to transfer the modality-level complementarities for efficient knowledge transfer and importantly do not interfere with the learning of unimodal knowledge. To achieve this, we propose a novel Dimension-decoupled Unidirectional Contrastive (DUC) loss. Due to factors such as over-parameterization and implicit regularization [4, 131], deep networks tend to learn low-rank and redundant features, which motivates us to **compensate the ineffective information present in features with the effective cross-modal complementary information**.

First, we need to perform dimension separation to specify the effective and ineffective dimensions for each modality. We define the effective dimensions as dimensions with better discriminative knowledge. Therefore, we devise the dimension-wise prediction to evaluate the discrimination for each modality. With all the features from modality i, we can obtain the feature centroid of each class as:

$$\bar{\boldsymbol{h}}_k^i = \frac{1}{N_k}\sum_{j=1}^{N} \mathcal{I}\left\{y_j = k\right\} \boldsymbol{h}_j^i,\ \bar{\boldsymbol{h}}_k^i = \left[\bar{h}_{k,1}^i, \bar{h}_{k,2}^i, ..., \bar{h}_{k,d}^i\right]^T \tag{3.3}$$

where N is the number of all samples and N_k is the number of samples belong to k-th class. And then, we can make dimension-wise evaluation by comparing the distance for each dimension with its dimensional centroid:

$$r_m^i = \frac{1}{N}\sum_{j=1}^{N} \mathcal{I}\left\{\arg\min_k d\left(h_{j,m}^i, \bar{h}_{k,m}^i\right) = y_j\right\}, m \in [d] \tag{3.4}$$

$d\left(\cdot,\cdot\right)$ is the distance function (Euclidean distance here). r_m^i can be used to assess the effectiveness of dimension m of modality i. Larger value indicates higher effectiveness on classification. Hence, the dimension separation principle is that the effective dimensions are represented with dimensions whose dimension-wise evaluation is greater than the mean value:

$$\begin{cases} r_m^i > \bar{r}^i & m\ is\ effective \\ r_m^i < \bar{r}^i & m\ is\ ineffective \end{cases} \tag{3.5}$$

where $\bar{r}^i = \frac{1}{d}\sum_{m=1}^{d} r_m^i$. Through this way, the feature dimensions of each modality are divided into effective group $d_e^i = \{m | r_m^i > \bar{r}^i\}$ and ineffective group $d_{ne}^i = \{m | r_m^i < \bar{r}^i\}$. The dimension separation is operated after some warm-up epochs.

Due to the heterogeneity between modalities, they do not share all the effective dimensions. Hence, we then propose to transfer the effective information in modality 1 to the corresponding ineffective dimensions in modality 2 and vice verse, as shown in Figure 3.3. The knowledge transfer is performed by our proposed DUC loss:

$$\begin{aligned}
\mathcal{L}_{DUC}^1 &= \mathbb{E}_{\left(x_i^1, x_i^2\right)}\left[-\log \frac{\exp\left(-d\left(\tilde{\boldsymbol{h}}_i^1, \tilde{\boldsymbol{h}}_i^2\right)/T\right)}{\sum_j \exp\left(-d\left(\tilde{\boldsymbol{h}}_i^1, \tilde{\boldsymbol{h}}_j^2\right)/T\right)}\right] \\
\mathcal{L}_{DUC}^2 &= \mathbb{E}_{\left(x_i^1, x_i^2\right)}\left[-\log \frac{\exp\left(-d\left(\hat{\boldsymbol{h}}_i^1, \hat{\boldsymbol{h}}_i^2\right)/T\right)}{\sum_j \exp\left(-d\left(\hat{\boldsymbol{h}}_j^1, \hat{\boldsymbol{h}}_i^2\right)/T\right)}\right]
\end{aligned} \tag{3.6}$$

where $\tilde{\boldsymbol{h}}_i^1 = \left[h_{i,m}^1 | m \in d_{ne}^1 \cap d_e^2\right]$, $\tilde{\boldsymbol{h}}_i^2 = \left[h_{i,m}^2 | m \in d_{ne}^1 \cap d_e^2\right]$, $\hat{\boldsymbol{h}}_i^1 = \left[h_{i,m}^1 | m \in d_e^1 \cap d_{ne}^2\right]$ and $\hat{\boldsymbol{h}}_i^2 = \left[h_{i,m}^2 | m \in d_e^1 \cap d_{ne}^2\right]$. T is the temperature. Notably, **the features of $\tilde{\boldsymbol{h}}_i^2$ and $\hat{\boldsymbol{h}}_i^1$ do not pass gradient backward**, which means we only allow the ineffective dimensions of modality 1 (2) to learn toward the corresponding effective dimensions of modality 2 (1), and do not update the effective dimensions of modality 2 (1) with DUC to prevent damage on the unimodal learning process. Hence, we let the complementary knowledge between modalities transfer unidirectionally and use the integrated knowledge for prediction to enable cross-modal interaction.

The final loss for modality i can be calculated as:

$$\mathcal{L}^i = \mathcal{L}_{CE}^i + \lambda_s \mathcal{L}_{CE}^{Si} + \lambda_D \mathcal{L}_{DUC}^i \tag{3.7}$$

Interaction during co-prediction. The above training process does not directly utilize the multimodal data for completing tasks, therefore, in this stage we enable the interaction during the co-prediction process via training a fusion module with multimodal objective Eq. 3.1 while fixing the learned encoders.

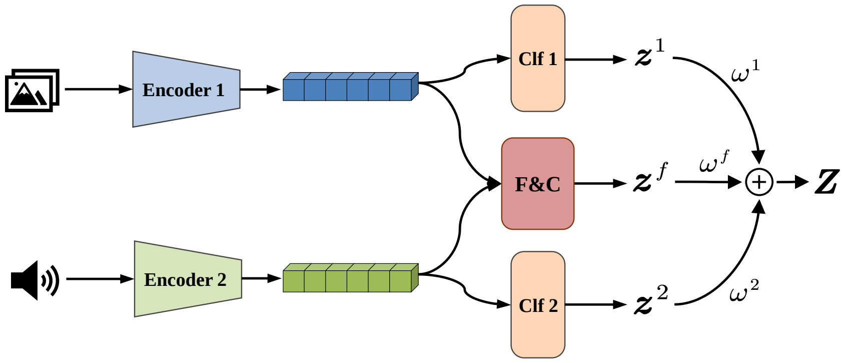


Figure 3.4: During inference, the logit weighting is utilized on instance level.

3.4.2 Instance-level Weighting

In the training stage, we exploit the modality-level complementary information through DUC loss. However, the complementary capacities of the different modalities may also vary in different sample pairs [109]. Therefore, we propose a certainty-aware logit weighting strategy during inference to utilize the instance-level complementarities comprehensively, as demonstrated in Figure 3.4. We use the absolute certainty to evaluate the j-th instance reliability for each modality and their fusion:

$$c_j^i = \max_k softmax\left(\boldsymbol{z}_j^i\right)_k, \ i \in \{1, 2, f\}, \ k \in [K]. \tag{3.8}$$

superscript f denotes the output of fusion module. Then, the final output is:

$$\begin{aligned} \boldsymbol{Z}_j &= w_j^1 \boldsymbol{z}_j^1 + w_j^f \boldsymbol{z}_j^f + w_j^2 \boldsymbol{z}_j^2 \\ w_j^i &= \frac{\exp\left(c_j^i/T\right)}{\exp\left(c_j^1/T\right) + \exp\left(c_j^f/T\right) + \exp\left(c_j^2/T\right)} \end{aligned} \tag{3.9}$$

where more reliable modalities are assigned with higher weights.

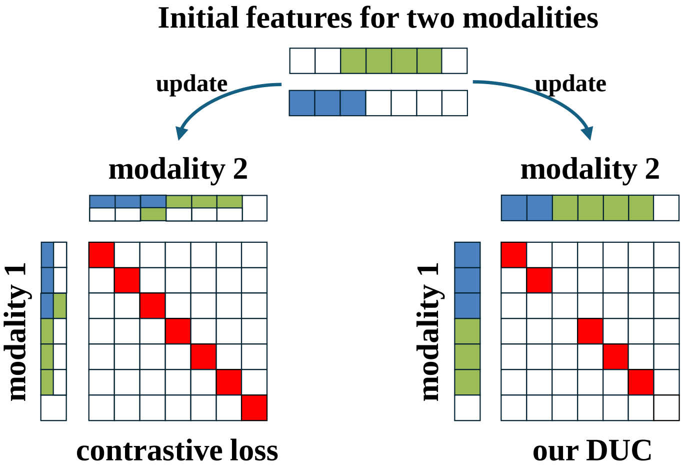


Figure 3.5: Traditional contrastive loss is hard, aligning all the dimensions bidirectionally. Our DUC loss is soft, performing on part of dimensions and only transferring complementarities. Blue and green colors denote effective dimensions and white means ineffective dimension. Red color represents alignment between corresponding dimensions.

3.4.3 Comparison with MCRL Loss

Previous multimodal contrastive loss [87] pays attention to searching for the semantic alignment between modalities, hence, the learning strength is bidirectional on the whole dimensions, i.e. the positive samples of two modalities move toward each other. Nevertheless, the alignment objective is too 'hard' that may lead to information loss, since there may be noise in part of the dimensions for specific modalities and complete alignment would partially preserve the noise, as illustrated in Figure 3.5. In contrast, our DUC loss is not intended to perform semantic alignment, but rather cross-modal transfer of complementary knowledge. Therefore, we decouple the feature dimensions and perform a unidirectional cross-modal knowledge transfer to enhance the dimensions with less informative knowledge while retaining effective information unique to the current modality. It can be seen that our DUC is more 'soft', and the dimensions in $d_e^1 \cap d_e^2$ are not required to align with each other, preserving the specific characteristics of each modality.

Table 3.2: Comparative analysis of different methods on CREMA-D and AVE. The metric is the top-1 accuracy (%). The best performance is in **bold**, and the second best is underlined.

Dataset	CREMA-D [9]			AVE [99]		
Method	Audio	Visual	Multi	Audio	Visual	Multi
Uni1	65.59	-	-	**66.42**	-	-
Uni2	-	78.49	-	-	46.02	-
Joint training	61.96	38.58	70.83	63.93	24.63	69.65
MSES [35]	62.50	37.90	70.43	63.93	24.63	69.65
MSLR [126]	63.04	41.13	71.51	61.19	24.63	68.91
OGM-GE [85]	61.29	39.27	71.14	62.45	27.39	69.12
PMR [30]	63.04	71.24	75.54	63.18	35.57	70.89
UMT [25]	65.46	75.94	77.42	65.42	42.29	73.88
MM Clf	65.59	78.49	78.09	66.42	46.02	72.39
Preds Avg	65.59	78.49	82.66	66.42	46.02	69.40
Ours	**66.67**	**78.90**	**83.74**	64.18	**49.25**	**75.37**

3.5 EXPERIMENTS

3.5.1 Dataset

We use four different multimodal datasets, *i.e.*, CREMA-D [9], AVE [99], UCF101 [94], and ModelNet40. CREMA-D is an audio-visual dataset for researching emotion recognition, comprising facial and vocal emotional expressions. Emotions are categorized into 6 types: happy, sad, angry, fear, disgust, and neutral. The dataset consists of 7442 segments, randomly divided into 6698 samples for training and 744 samples for testing. AVE is an audio-visual video dataset designed for audio-visual event localization, encompassing 28 event classes and 4,143 10-second videos. It includes both auditory and visual tracks along with secondary annotations. All videos are collected from YouTube. In our experiments, we extract frames from event-localized video segments and capture audio clips within the same segment, constructing a labeled multimodal classification dataset as in [30]. UCF101 is a dataset for action recognition comprising real action videos with 101 action categories, collected from YouTube.

Table 3.3: Comparative analysis of different methods on UCF101 and ModelNet40. The metric is the top-1 accuracy (%). The best performance is in **bold**, and the second best is underlined.

Dataset	UCF101 [94]			ModelNet40		
Method	Flow	Image	Multi	Front	Rear	Multi
Uni1	55.09	-	-	89.63	-	-
Uni2	-	42.96	-	-	88.70	-
Joint training	33.78	37.54	51.92	85.98	81.81	89.63
MSES [35]	33.99	37.19	51.76	85.98	81.81	89.63
MSLR [126]	33.44	37.77	52.60	86.22	82.17	89.59
OGM-GE [85]	-	-	52.92	-	-	89.30
PMR [30]	-	-	-	-	-	-
UMT [25]	55.41	45.15	61.51	88.33	87.76	90.80
MM Clf	55.09	42.96	60.67	89.63	88.70	90.19
Preds Avg	55.09	42.96	64.43	89.63	88.70	90.92
Ours	**58.52**	**48.59**	**65.79**	**89.83**	**88.74**	**90.92**

We treat the optical flow and images of the videos as two separate modalities. The dataset consists of 13,320 videos, with 9,537 used for training and 3,783 for testing. ModelNet40 is one of the Princeton ModelNet datasets [113] with 3D objects of 40 categories, consisting of 9,843 training samples and 2,468 testing samples. Following [112], we treat the front view and the rear view as two modalities.

3.5.2 Experimental Settings

For the above four datasets, we used ResNet18 [44] as the backbone encoder network, mapping input data into 512-dimensional vectors. For the input data, for the CREMA-D and AVE datasets, audio modality data was transformed into spectrograms of size 257×1,004, and visual modality data consisted of 3(4 frames for AVE) randomly selected frames from 10-frame video clips, with image size of 224×224. For the UCF101 dataset, we randomly sampled contiguous 10-frame segments from videos during training, while testing, we sampled 10-frame segments from the middle of the videos. Optical flow modality data was of size 20×224×224, and visual

Table 3.4: The ablation study on CREMA-D and AVE.

TS	S-Clf	DUC	LW	CREMA-D			AVE		
				Audio	Visual	Multi	Audio	Visual	Multi
				61.96	38.58	70.83	63.93	24.63	69.65
✓				65.59	78.49	78.09	**66.42**	46.02	72.39
✓	✓			66.26	79.70	79.70	64.43	44.78	72.14
✓	✓	✓		66.67	78.90	82.80	64.18	49.25	72.89
✓	✓	✓	✓	**66.67**	**78.90**	**83.74**	64.18	**49.25**	**75.37**

modality data consisted of randomly sampled 1 frame. For the ModelNet40 dataset, we utilized front and back views as two modalities. For all visual modalities, we applied random cropping and random horizontal flipping as data augmentation during training; we resized images to 224×224 without any augmentation during testing. We trained all models with a batch size of 16, using SGD optimizer with momentum of 0.9 and weight decay of 1e-4, for a total of 150 epochs, with initial learning rate of 1e-3 decaying to 1e-4 after 70 epochs. For the second stage of our method, we trained for 20 epochs, with initial learning rate of 1e-3 decaying to 1e-4 after 10 epochs. All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU.

3.5.3 The Effectiveness of DI-MML

We compare DI-MML with various baselines and validate the effectiveness of our method.

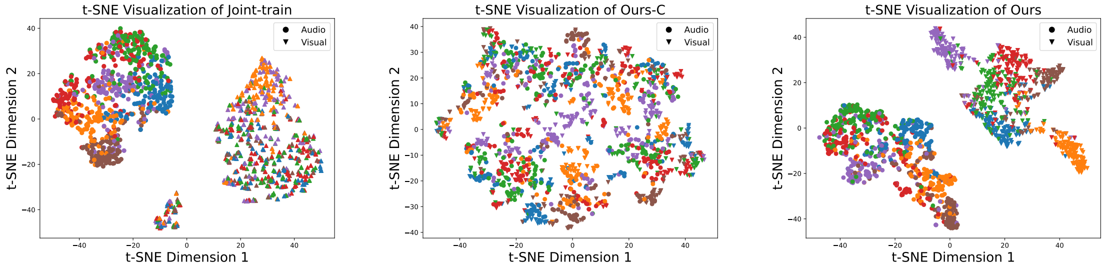
Comparison with other baselines. The compared methods are divided into two groups: with and without the uniform objective for encoder training. Only the MM Clf, Preds Avg and our DI-MML do not utilize the uniform objective. The results are shown in Tables 3.2 and 3.3, we not only report the multimodal performance and also the unimodal accuracy. To ensure the fairness of the comparison, we fix the parameters of their unimodal encoder networks after multimodal training, and evaluate their unimodal performance by training a classifier independently. It can be that the

Table 3.5: The performance comparison with various contrastive losses.

Dataset	CREMA-D			AVE		
Method	A	V	Multi	A	V	Multi
w/o DUC	66.26	79.70	83.47	**64.43**	44.78	74.13
Our-C	65.73	79.17	81.72	63.18	46.77	71.39
Our-DBC	65.99	**79.84**	82.12	63.18	**49.50**	73.13
Ours	**66.67**	78.90	**83.74**	64.18	49.25	**75.37**

methods with the uniform objective (joint training, MSES, MSLR, OGM-GE, PMR and UMT) are all suffered from severe modality competition as their unimodal performance is generally lower than the best unimodal training counterpart, especially on Visual in CREMA-D and AVE, Flow in UCF101 and Rear in ModelNet40. MSES, MSLR, OGM-GE and PMR regulate the learning progress of modalities by adjusting the learning rates or gradients of different modalities, which alleviates modality competition to some extent, but they are difficult to completely eradicate it. UMT maintains the performance of the different modalities better, but it requires pretrained unimodal models for distillation, which is expensive and impractical. In contrast, our method completely avoid the modality competition, resulting in comparable or even the best unimodal performance (improved by up to 3.11% and 3.44% on Flow and Image of UCF101) and the best multimodal performance (improved by up to 6.32% on CREMA-D) on all four datasets. Besides, we do not require additional computational cost for encoder training. Compared with MM Clf and Preds Avg, our DI-MML enables cross-modal interaction and complementary knowledge transfer during the encoder training. Therefore, our method can achieve both better multimodal and unimodal performance on these datasets. These results show that our approach is indeed competition-free and the proposed cross-modal interactions are effective.

Ablation study. There are four main components in our method: two-stage training scheme (TS), shared classifier (S-Clf), dimension-decoupled unidirectional contrastive loss (DUC), and logit weighting (LW). Here, we perform an ablation study to explore the influence of various combinations of these components. The experiments are con-



(a) Joint training (b) Our-C (c) Ours

Figure 3.6: The t-SNE feature visualization of each modality on CREMA-D. Different colors denote different classes.

Table 3.6: The number of effective dimensions for each modality on three datasets. ‘Overlap’ denotes $|d_e^1 \cap d_e^2|$. The results are obtained from the model after warmup epochs.

	CREMA-D	AVE	UCF101
Audio/Flow eff	259	258	246
Visual/Image eff	262	291	249
Overlap	156	142	138

ducted on CREMA-D and AVE. As demonstrated in Table 3.4, applying TS denotes the MM Clf method, which is better than Joint training because there is no modality competition. The shared classifier can align a feature space for different modalities and achieve considerable improvement on CREMA-D. The DUC loss facilitates cross-modal interaction and knowledge transfer, helping to achieve complementary knowledge utilisation at the modality level. Similarly, LW enables complementary knowledge integration at the instance level, both of them are important for multimodal performance enhancement. As discussed above, the four components are all essential in our method.

Analysis on DUC loss. The DUC loss is the central technique in our method to enhance the cross-modal interaction during the encoder training stage. In subsection 3.4.3, we compare the differences between DUC and traditional multimodal contrastive learning loss in terms of aim and formality. Here, we give more experimen-

Table 3.7: The performance of effective and ineffective dimensions of each modality.

Dataset	CREMA-D			AVE		
Modality	all	eff	ineff	all	eff	ineff
Audio	58.60	54.71	31.59	59.70	50.25	43.03
Visual	46.37	31.99	23.79	25.12	21.64	18.91

Table 3.8: The performance of different methods for evaluating the effectiveness of each dimension.

Dataset	CREMA-D	AVE
Method	Multi	Multi
Joint training	70.83	69.65
L2-norm	83.60	73.17
Shapley value	81.58	**75.37**
Dimension-wise prediction	**83.74**	**75.37**

tal results to show the superiority of our method. The results are shown in Table 3.5, where '-C' denotes replacing our DUC loss with traditional multimodal contrastive loss while '-DBC' means dimension-decoupled bidirectional contrastive loss, *i.e.*, $\tilde{\boldsymbol{h}}_i^2$ and $\hat{\boldsymbol{h}}_i^1$ are not detached in Eq. 3.6, suggesting that $\tilde{\boldsymbol{h}}_i^1$ and $\tilde{\boldsymbol{h}}_i^2$ ($\hat{\boldsymbol{h}}_i^1$ and $\hat{\boldsymbol{h}}_i^2$) move toward each other as traditional contrastive loss. It is clear that using traditional contrastive loss performs worst as it does not consider retaining the modality-wise complementary information. Applying DBC achieves improvement since the it does not affect the learning of effective dimensions shared by modalities (*i.e.*, $d_e^1 \cap d_e^2$). However, the noise information in the ineffective dimensions is preserved as illustrated in Figure 3.5. Our DUC loss both preserves the complementary knowledge of each modality and facilitates inter-modal cooperation through knowledge transfer, resulting in the best multimodal results. In Figure 3.6, we demonstrate the t-SNE [100] feature visualization for each modality on CREMA-D. Figure 3.6(a) showcases that the there are no clear decision boundaries for visual features, consistent with its poor performance. As shown in Figure 3.6(b), although applying contrastive loss in our method compensates for the gap between different modalities in feature space, the noise in visual modality is also preserved to some extent, leading to worse multimodal

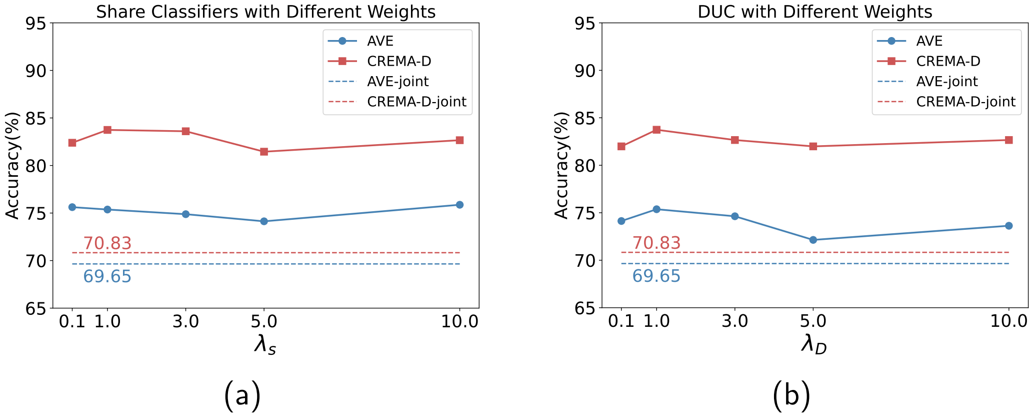


Figure 3.7: Comparison with different values of λ_s and λ_D.

performance. With the optimization from our method as shown in Figure 3.6(c), the features of both modalities are more clearly clustered, besides, share a more similar distributional structure.

Analysis on dimension separation. In this section, we perform the dimension separation to divide dimensions into effective and ineffective parts. The separation results are displayed in Table 3.6. The effective dimensions for both modalities take up about half or more (feature is a 512-dimensional vector), and their overlap also accounts for only about half of effective dimensions, indicating that there are enough dimensions to ensure the effectiveness for cross-modal knowledge transfer. The performance of corresponding dimension sets of effectiveness and ineffectiveness is shown in Table 3.7. When we evaluate the performance of effective dimension set, the values of ineffective dimensions are set to 0 and vice verse. The performance of effective dimensions is much better than that of ineffective dimensions, indicating that our dimension separation scheme is reasonable and effective.

3.5.4 Robustness Validation

Effective dimension evaluation. In this section, we devise the dimension-wise prediction as in Eq. 3.4 to evaluate the effectiveness of each dimension. Here, we compare our method with two other evaluation metrics: L2-norm and Shapley Value. According to [86], the L2-norm of the features gives an indication of their information content, thus it can be used as a metric to measure the effectiveness of each dimension. And shapley value can also be used to identify important features (dimensions here). As depicted in Table 3.8, our proposed framework has significant enhancements with any evaluation method, showing the robustness of our method. Besides, among the three methods, our dimension-wise prediction performs the best on different datasets, indicating its validity for evaluating the dimensionally discriminative information.

Hyperparameter sensitivity. In the calibration of our DI-MML, we encounter two hyperparameters to determine: λ_s and λ_D in Eq. 4.12, determining the strength for feature space alignment and cross-modal knowledge transfer respectively. We explore the effects of them as illustrated in Figure 4.5. It is clear that the performance on DI-MML is marginally affected by λ_s and λ_D, suggesting the insensitivity of our method to hyperparameters. Despite some fluctuations in performance with hyperparameters, it still demonstrates excellent effectiveness (consistently better than joint training). We select $\lambda_s = 1$ and $\lambda_D = 1$ for the best results.

3.6 Remarks

In this chapter, we analyze the multimodal joint training and argue that the modality competition problem comes from the uniform learning objective for different modalities. Therefore, we propose to train multimodel encoders separately to avoid modality competition. To facilitate the feature space alignment and cross-modal interaction, we devise a shared classifier and the dimension-decoupled unidirectional contrastive loss (DUC) to achieve modality-level complementarities utilization. And then, the learned encoders are frozen and a fusion module is updated for interaction during co-prediction. Considering the reliability differences on various sample pairs, we further propose the certainty-aware logit weighting strategy to exploit instance-level complementarities comprehensively. Through extensive experiments, our DI-MML outperforms all competing methods in four datasets. We also showcase that our method can further promote the unimodal performance instead of inhibiting them. In the future, we can investigate other types of cross-modal interactions and focus on multimodal tasks such as detection or generation instead of only classification. Besides, identifying the specific semantics in each dimension may be helpful to further evaluate the informative dimensions.

Chapter 4

Cross-modal Representation Flattening for Multi-modal Domain Generalization

Multi-modal domain generalization (MMDG) requires that models trained on multi-modal source domains can generalize to unseen target distributions with the same modality set. Sharpness-aware minimization (SAM) is an effective technique for traditional uni-modal domain generalization (DG), however, with limited improvement in MMDG. In this section, we identify that *modality competition* and *discrepant uni-modal flatness* are two main factors that restrict multi-modal generalization. To overcome these challenges, we propose to construct consistent flat loss regions and enhance knowledge exploitation for each modality via cross-modal knowledge transfer. Firstly, we turn to the optimization on representation-space loss landscapes instead of traditional parameter space, which allows us to build connections between modalities directly. Then, we introduce a novel method to flatten the high-loss region between minima from different modalities by interpolating mixed multi-modal representations. We implement this method by distilling and optimizing generalizable interpolated representations and assigning distinct weights for each modality considering their divergent generalization capabilities. Extensive experiments are performed on two benchmark datasets, EPIC-Kitchens and Human-Animal-Cartoon (HAC), with various modality combinations, demonstrating the effectiveness of our method under multi-source and single-source settings.

4.1 Introduction

Domain generalization (DG) aims to equip models with the ability to perform robustly across unseen domains when trained only on several source domains, thereby enhancing their adaptability and utility in real-world scenarios, such as autonomous driving [91, 17], medical health [63, 73], person re-identification [6, 81] and brain-computer interface [78, 42]. The statistical distribution gap between target domains (where a model is applied) and source domains (where the model was trained) is defined as domain shift. Methods on how to deal with domain shift have been extensively proposed in the literature, including domain alignment [108], meta-learning [7, 69], data augmentation [136, 135] and ensemble learning [10]. Despite the remarkable achievements of DG in recent years, most of research still focuses on uni-modal data. The emergence of various multi-modal datasets and the requirement to complete a variety of multi-modal tasks highlight the need to address multi-modal domain generalization (MMDG) problems.

Due to the complementary information that exists between modalities, MMDG aims to exploit generalization capabilities from each modality simultaneously. According to [53], the generalization capability of deep neural networks (DNNs) is closely related to their flatness of minima on loss landscape (as shown in Fig. 4.1 (a)), which motivates penalizing sharpness [11] and rewarding flatness [51]. Sharpness-aware minimization (SAM) [34] and its variants [10, 133] have been proposed to seek flatter minima and achieve better generalization across domains. Despite their success on uni-modal

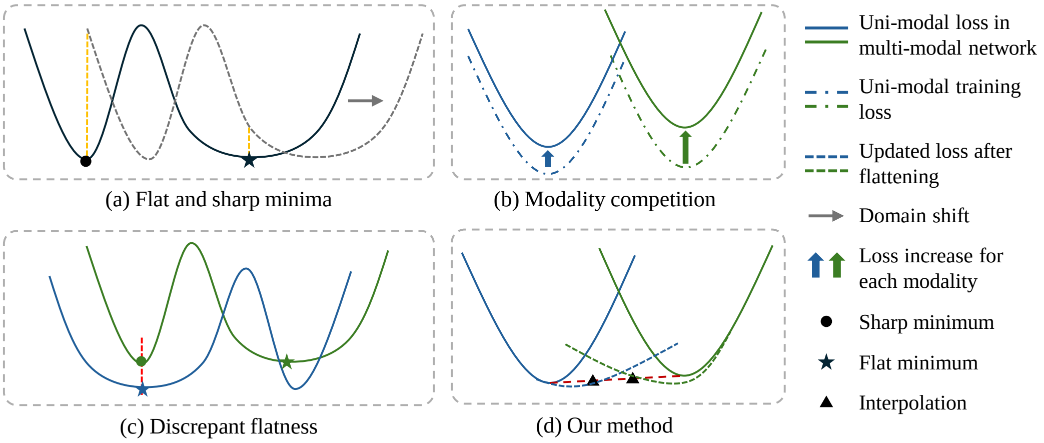


Figure 4.1: (a) Flat minima on loss landscape generalize better than sharp minima with domain shift. (b) Multi-modal joint training leads to larger loss for each modality compared with independent uni-modal training. (c) The flat minima between modalities are usually inconsistent, making it hard to obtain flat minima for each modality simultaneously in a multi-modal network. (d) We optimize the cross-modal interpolations on representation-space loss landscape to get consistent flat region.

scenarios, in this section, we argue that they are not compatible well in MMDG since the distinct properties between modalities pose two challenges (more details can be found in Sec.4.4). (1) **Modality competition**: according to [49], multiple modalities will compete with each other during joint training, leading to inadequate knowledge exploitation for each modality [29, 28], i.e, larger minima of loss as shown in Fig. 4.1 (b), and consequently worse generalization. (2) **Discrepant uni-modal flatness**: the generalization gap between modalities makes it hard to find their flat minima simultaneously, resulting in multi-modal networks incapable of utilizing generalization capabilities from all modalities, as illustrated in Fig. 4.1 (c). Hence, existing methods can not fully exploit the generalization potential of each modality, which inevitably leads to sub-optimal solutions for MMDG.

To overcome these challenges, we propose to construct consistent flat loss regions and enhance knowledge exploitation for each modality via cross-modal knowledge transfer. Traditional SAM-based methods are analyzed on parameter space. However, due to the heterogeneity between modalities, their parameter spaces could be extremely different (e.g., different model structures and parameter numbers), making it challenging to represent their correlation. Instead, we turn to optimization on representation-space loss landscape [138] as representations of different modalities can be mapped into a shared space, so that we can build their connections directly. Based on this, we propose a novel **C**ross-**M**odal **R**epresentation **F**lattening (CMRF) method to achieve consistent representation flat minima. As shown in Fig. 4.1 (d), we construct the interpolations by mixing paired multi-modal representations and then optimize them to flatten the high-loss regions between minima from different modalities. Specifically, we obtain more stable and generalizable cross-modal interpolations from moving averaged teacher model and then employ feature distillation to regularize the learning of each modality. The interpolations between modalities bring their flat regions closer, alleviating their flatness discrepancy. Moreover, the cross-modal knowledge transfer also helps to promote each modality and alleviate their competition. Our contributions can be summarized as:

- To the best of our knowledge, we are the first to extend the uni-modal flatness analysis to MMDG, and empirically attribute the reasons for limited MMDG performance to two problems: modality competition and discrepant uni-modal flatness.

- We construct shared representation space instead of parameter space to build connections between modalities directly and propose to flatten high-loss representation regions between modalities by interpolating mixed multi-modal representations and performing knowledge distillation to regularize the learning of each modality.

- Extensive experiments verify the effectiveness and superiority of our framework on two benchmark datasets of EPIC-Kitchens and Human-Animal-Cartoon (HAC) under various modalities combinations on both multi- and single-source MMDG.

4.2 Related Work

4.2.1 Flat Minimum of Loss Landscape for DG

Domain generalization refers to the ability of models to perform well on new, unseen domains that are dissimilar with domains they were trained on. Numerous methods have been proposed to tackle the domain shift, while one type among them is to search for flat minima in loss landscapes [34, 132, 133]. Jiang *et al.* [53] conducted comprehensive measures and found that a sharpness-based measure has highest correlation with generalization. Based on that, Foret *et al.* [34] proposed sharpness-aware minimization (SAM) to seek parameters that lie in neighborhoods with uniformly low loss via perturbed gradients, while Wang *et al.* [105] further proposed to align the gradient directions between the empirical risk and the perturbed loss. Moreover, average weights during training has also shown to yield flatter minima [51], which motivates more elegant average methods such as SWAD [10] and EoA [5]. In this section, we try to optimize consistent flat minima for different modalities in representation-space loss landscapes instead of traditional parameter space.

4.2.2 Multi-modal DG

Although uni-modal DG has been extensively studied in recent years, the research on MMDG is severely insufficient, while only few works have been done. Planamente *et al.* [86] proposed RNA-Net to balance audio and video feature norms via a relative norm alignment loss. Dong *et al.* [23] proposed a unified framework to achieve domain generalization in various multimodal scenarios including multi-source, unisource, and modality missing DG. In this section, we extend the unimodal flatness analysis to MMDG and address two particular problems in multi-modal scenarios.

4.2.3 Mixup

Mixup [130] is a data augmentation technique introduced to improve the generalization performance of models. Traditional mixup and its variant CutMix [129] are performed on input data, while Verma *et al.* [101] further introduced Manifold Mixup that mixes the representations in each layer to produce smoother decision boundaries. However, Manifold Mixup and its variants [74, 121] are designed for uni-modal data, and only few works are on multi-modal scenarios [32, 83]. STEMM [32] aims to align speech and text features by mixing them, but is limited with its architecture-specific design. Oh *et al.* [83] introduced m^2-Mix aiming at generating hard negative samples by mixing image and text embeddings to fine-tuning CLIP. Compared with them, our mixed multi-modal representations has no architecture restrictions and are used as teacher signals to guide various modalities to learn consistent flat minima.

4.3 Preliminaries of MMDG

We follow the definition of multi-modal domain generalization problem as in [23]. In MMDG, we are given D source domains for training $\mathcal{D}_{train} = \{\mathcal{D}^i | i = 1, \cdots, D\}$, where $\mathcal{D}^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i} \sim P_{XY}^i$ denotes the i-th domain with n_i data instances sampled from a joint distribution of input samples and output labels P_{XY}^i. X and Y represent the corresponding random variables. Each input instance $\mathbf{x}_j^i = \left\{ \left(\mathbf{x}_j^i\right)_k | k = 1, \cdots, M \right\} \in \mathbf{X}$ consists of M different modalities and $y_j^i \in \mathcal{Y} \subset \mathbb{R}$ denotes corresponding label, where $\mathbf{X}$ and $\mathcal{Y}$ represent input and output space. The joint distributions in $\mathcal{D}_{train}$ are different from each other: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq D$. Now, with an unseen test domain $\mathcal{D}_{test}$ with M modalities that cannot be accessed during training and $P_{XY}^{test} \neq P_{XY}^i$ for $i \in \{1, \cdots, D\}$, the goal of MMDG is to learn a robust and generalizable predictive function $f : \mathbf{X} \rightarrow \mathcal{Y}$ based on D training domains to achieve a minimum prediction error on $\mathcal{D}_{test}$:

$$\min_{f} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_{test}} \left[\ell\left(f\left(\mathbf{x}\right), y\right)\right] \tag{4.1}$$

where $\mathbb{E}$ is the expectation and $\ell\left(\cdot,\cdot\right)$ is the loss function, e.g., cross-entropy loss for multi-modal classification tasks. In this section, we use $\theta = \{\theta_1, \cdots, \theta_M\}$ to denote the parameters of the neural network f, where θ_i indicates the parameters for i-th modality. Therefore, the training loss over all training domains $\mathcal{D}_{train}$ is defined as follows:

$$\mathcal{L}\left(\theta; \mathcal{D}_{train}\right) = \frac{1}{\sum_{i=1}^{D} n_i} \sum_{i=1}^{D} \sum_{j=1}^{n_i} \ell\left(f\left(\mathbf{x}_j^i; \theta\right), y_j^i\right) \tag{4.2}$$

The empirical risk minimization (ERM) of Eq. 4.2 tends to converge to sharp minima and SAM [34] is proposed to seek flatter minima on loss landscape with the following optimization:

$$\min_{\theta} \mathcal{L}\left(\theta + \hat{\epsilon}; \mathcal{D}_{train}\right), \text{ where } \hat{\epsilon} \triangleq \rho \frac{\nabla\mathcal{L}\left(\theta; \mathcal{D}_{train}\right)}{\|\mathcal{L}\left(\theta; \mathcal{D}_{train}\right)\|}. \tag{4.3}$$

where ρ is a predefined constant controlling the radius of the neighborhood.

4.4 MMDG Analysis

MMDG aims to comprehensively exploit the generalization capabilities from each modality to learn more robust and generalized models. However, the generalization behavior of each modality in multi-modal networks has not been well explored. Here, we analyze the behavior of each modality and find the challenges for generalizable multi-modal networks.

Modality competition leads to larger minima. As demonstrated in Tab. 4.1, we compare naive joint training and SAM about their uni- and multi-modal performance.

Table 4.1: MMDG analysis on EPIC-Kitchens and HAC with video and audio data. 'Base' denotes the naive multi-modal joint training without any domain generalization strategies. 'Uni-video' and 'Uni-audio' means training only with uni-modal data. 'Video', 'Audio' and 'Video-Audio' denote testing with uni-modal and multi-modal data. Results are averaged by using each domain as target.

	EPIC-Kitchens			**HAC**		
	Video	Audio	Video-Audio	Video	Audio	Video-Audio
Uni-video	58.73	-	-	68.07	-	-
Uni-audio	-	40.04	-	-	32.81	-
Uni-video-SAM	**61.68**	-	-	<u>69.58</u>	-	-
Uni-audio-SAM	-	<u>42.65</u>	-	-	**35.84**	-
Base	56.65	38.62	59.63	67.60	31.24	63.11
SAM	58.80	37.77	<u>61.19</u>	68.46	31.56	<u>64.72</u>
CMRF (ours)	<u>60.66</u>	**43.13**	**63.91**	**70.54**	<u>34.86</u>	**71.91**

SAM can clearly improve generalization on both uni-modal and multi-modal training. However, the uni-modal generalization from multi-modal trained network is worse than uni-modal trained network, whether or not SAM is applied (e.g, 56.65% vs. 58.73% without SAM and 58.80% vs. 61.68% with SAM on EPIC-Kitchens video). This phenomenon can be explained by modality competition [49, 31] that modalities in joint training compete with each other, making each modality under-explored. Our empirical results show that it not only degrades in-domain performance for each modality as discussed in [85, 30], but also weakens their out-of-domain generalization, resulting in larger minima of loss as shown in Fig. 4.1 (b).

Generalization gap results in discrepant uni-modal flatness. We observe that applying SAM can only improve generalization of better modality in multi-modal network but has marginal benefit or even harm on weak modality (e.g., video generalization is improved from 56.65% to 58.80% on EPIC-Kitchens while the number of audio drops from 38.62% to 37.77%). According to [30], the better modality will dominate multi-modal gradients. Hence, in Eq. 4.3, the gradient perturbation $\hat{\epsilon}$ in SAM could also be dominated by the better modality, which means this optimization on multi-modal network tends to search for flatter regions for modality with better generalization but ignores other weak modalities. This suggests that conventional uni-modal SAM-based methods cannot find the coexisting flat minima for each modality due to their generalization gap, leading to discrepant flatness and consequently under-utilization of generalization from all modalities, as shown in Fig. 4.1 (c). More results with other modality combinations can be found in Sec. 4.6.2.

4.5 Methodology

4.5.1 Cross-Modal Representation Flattening

Based on the analyses above, in this section, we aim to 1) accomplish consistent flat minima for all modalities in multi-modal network and 2) alleviate the competition between modalities to utilize their generalization comprehensively. Considering the correlation and complementary information between modalities, we propose to leverage cross-modal knowledge transfer to enhance MMDG.

Representation-space loss landscape. Previous analysis of loss landscapes usually happens on parameter space [133, 58]. However, the network structures and sizes for different modalities are commonly different, leading to disparate parameter spaces. This makes it difficult to catch correlations between modalities and produce consistent flat loss regions in parameter space. Inspired by [138] that introduces representation-space loss landscape, we turn to analyze loss landscapes of different modalities in representation space. Specifically, given a data point $\mathbf{x}_j^i = \left\{ \left(\mathbf{x}_j^i\right)_k | k = 1, \cdots, M \right\}$, feature extractors are usually applied to transform input data into features with different dimensions:

$$\left(\boldsymbol{h}_j^i\right)_k = g_k\left(\left(\mathbf{x}_j^i\right)_k\right) \subset \mathbb{R}^{d_k} \tag{4.4}$$

where g_k is feature extractor for k-th modality, d_k is feature dimension size and $\exists k \neq l, d_k \neq d_l$. In this section, we use a projector $Proj_k(\cdot)$ for k-th modality

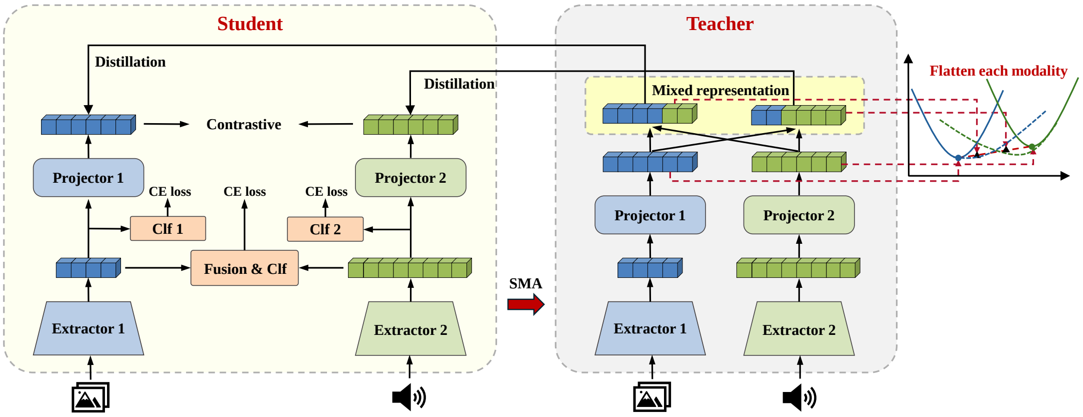


Figure 4.2: The overall framework of our method. The projectors map features with different dimensions to the same representation space. The teacher model is moving averaged from online model and generates cross-modal mixed representations as interpolations to distill the student representations. Uni-modal classifier is used to lower the loss of distilled features for each modality and a contrastive loss aims to alleviate gap between modalities. Only the online student model back propagates gradients. **The teacher model is used for evaluation finally.**

that maps its features into a shared representation space for all modalities with the same dimension d (omit superscript and subscript of domain and instance index for simplicity):

$$\boldsymbol{z}_k = Proj_k\left(\boldsymbol{h}_k\right) \subset \mathbb{R}^d,\ \ k \in \{1, \cdots, M\} \tag{4.5}$$

Given that each point in the representation space corresponds to a specific loss value, it is feasible to construct a landscape that maps each representation point to its associated loss value (e.g., horizontal axis indicates representation and vertical axis indicates loss in Fig. 4.1 (d)). After training, each representation extracted from each training sample can be viewed as a minimum. And we can judge whether a representation minimum is flat or sharp according to its neighboring loss distribution. In the shared representation loss landscape, we can build connections between different modalities directly.

Cross-modal representation interpolation. As discussed in Sec. 4.4, the discrepant uni-modal flatness severely impedes the utilization of generalization capability from each modality. The conclusion also applies to representation-space loss landscape since better modality still dominates gradients of representations, which optimizes weak modalities at sharp regions. Therefore, to obtain flat minima for various modalities simultaneously, we aim to flatten the high-loss regions between minima from different modalities. Given the paired multi-modal representations $\boldsymbol{z}_k$ and $\boldsymbol{z}_l$, $k \neq l$, we construct interpolated representations between them by cross-modal representation mixup:

$$\boldsymbol{z}_{k,l} = \delta \boldsymbol{z}_k + (1 - \delta)\, \boldsymbol{z}_l \tag{4.6}$$

where δ is mixing ratio. If the loss of mixed representations can be optimized to lower values, we would get a flatter region between modalities, as demonstrated in Fig. 4.1 (d). However, according to [101], directly optimization on mixed representations requires mixup at multiple eligible layers to be effective. It is impractical in multi-modal scenarios because representations of each layer for different modalities are generally at different scales, converting all them into a shared space is costly. In this section, we propose a simple yet effective method that distills the knowledge from mixed representations to each modality and then optimize the learned representations. Firstly, we perform simple moving average (SMA) [5] for the online updated network θ_k of each modality to establish the teacher network $\hat{\theta}_k^t$, which can produce more stable and generalizable representations:

$$\hat{\theta}_k^t = \begin{cases} \theta_k^t, & \text{if } t \leq t_0 \\ \frac{t-t_0}{t-t_0+1} \cdot \hat{\theta}_k^{t-1} + \frac{1}{t-t_0+1} \theta_k^t, & \text{otherwise} \end{cases} \tag{4.7}$$

where θ_k^t is the online model's state at iteration t of k-th modality. t_0 is the start iteration for SMA. Hence, the representation from teacher network is denoted as $\hat{\boldsymbol{z}}_k$ and the mixed representation of Eq. 4.6 should be rewritten as:

$$\hat{\boldsymbol{z}}_{k,l} = \delta \hat{\boldsymbol{z}}_k + (1 - \delta)\, \hat{\boldsymbol{z}}_l, \;\; \delta \sim Beta\,(\alpha, \alpha) \tag{4.8}$$

where α is a hyperparameter in Beta distribution. Considering the semantic gap between modalities, we let **interpolation closer** to k-th modality act as its teacher signal, so distillation loss should be:

$$\begin{cases} \mathcal{L}_{dis}^{k} = \frac{1}{M-1} \sum_{l=1,l\neq k}^{M} \left\| \boldsymbol{z}_k - \hat{\boldsymbol{z}}_{k,l} \right\|_2^2, & \delta > 0.5 \\ \mathcal{L}_{dis}^{l} = \frac{1}{M-1} \sum_{k=1,k\neq l}^{M} \left\| \boldsymbol{z}_l - \hat{\boldsymbol{z}}_{k,l} \right\|_2^2, & \delta < 0.5 \end{cases} \tag{4.9}$$

Then, we assign specific classifier for each modality before $Proj_k\left(\cdot\right)$ to online models and optimize the features by classification loss $\mathcal{L}_{cls}^{k}$. **The combination $\mathcal{L}_{dis}^{k} + \mathcal{L}_{cls}^{k}$ flattens the neighboring representation-space loss landscape of k-th modality to other modalities.** Further, we employ a multi-modal supervised contrastive loss on shared representation space, which can help to narrow the gap between modalities and make it conducive to flatten the region between them. For a random batch $\mathcal{B}$ with $M \times B$ uni-modal samples, we let i as the index of a uni-modal instance in the batch, and define $P\left(i\right)$ as the set of uni-modal samples that have the same label with i (except itself). The supervised contrastive loss can be written as (notably, subscript here does not denote modality index but the index of each sample):

$$\mathcal{L}_{con} = \sum_{i\in\mathcal{B}} -\frac{1}{\left|P\left(i\right)\right|} \sum_{p\in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p/\tau\right)}{\sum_{a\in\mathcal{B}\setminus\{i\}} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau\right)} \tag{4.10}$$

where $\tau \in \mathcal{R}^{+}$ is the temperature parameter.

4.5.2 Adaptive Weight

As demonstrated in Tab. 4.1, the generalization capabilities between modalities may have significant gaps, so we propose to assign stronger flattening weights to better modalities. We compare the uni-modal validation accuracy from teacher model (calculated by the moving averaged uni-modal classifier) as a rough estimate of the difference in generalization ability between modalities (the performance of different modalities on in-domain validation set can generally reflect their strength in generalization capability, as shown in Sec. 4.6.4). The distillation loss can be modified as:

$$\mathcal{L}_{dis}^{k} = \frac{1}{M-1} \sum_{l=1, l\neq k}^{M} \eta_{k,l} \left\| \boldsymbol{z}_k - \hat{\boldsymbol{z}}_{k,l} \right\|_2^2, \ \eta_{k,l} = \begin{cases} 1 & \hat{A}_k / \hat{A}_l > \mu \\ 0.5 & \hat{A}_k / \hat{A}_l \leq \mu \end{cases} \tag{4.11}$$

where $\hat{A}_k$ denotes the validation accuracy of k-th modality by teacher model, μ is a hyperparameter (default 1.2 in this section). In this way, the teacher signal with stronger generalization ability is applied with a larger distillation weight. Finally, we can get our final loss as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \sum_{k=1}^{M} \lambda_1 \mathcal{L}_{cls}^{k} + \sum_{k=1}^{M} \lambda_2 \mathcal{L}_{dis}^{k} + \lambda_3 \mathcal{L}_{con} \tag{4.12}$$

where $\mathcal{L}_{cls}$ is the multi-modal classification loss, and λ_1, λ_2 and λ_3 are hyperparameters to control the strength of each loss. Finally, we use teacher model for evaluation as it averages learned knowledge from student for better generalization.

4.6 Experiments

4.6.1 Experimental Setting

Dataset. We utilize two benchmark datasets: EPIC-Kitchens [21] and Human-Animal-Cartoon (HAC) [23]. Our experimental setup follows the protocols established for the EPIC-Kitchens dataset in [77] and for the HAC dataset in [23]. The EPIC-Kitchens dataset encompasses eight actions (‘put’, ‘take’, ‘open’, ‘close’, ‘wash’, ‘cut’, ‘mix’, and ‘pour’) captured across three different kitchens, forming three distinct domains: D1, D2, and D3. The HAC dataset comprises seven actions (‘sleeping’, ‘watching tv’, ‘eating’, ‘drinking’, ‘swimming’, ‘running’, and ‘opening door’) executed by humans (H), animals (A), and cartoon figures (C), resulting in three

Table 4.2: Multi-modal **multi-source** DG with different modalities on EPIC-Kitchens and HAC datasets. The best is in **bold**, and the second best is underlined.

	Modality			**EPIC-Kitchens**				**HAC**			
Method	Video	Audio	Flow	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*	A, C → H	H, C → A	H, A → C	*Avg*
Base	✓	✓		54.94	62.26	61.70	59.63	69.92	69.32	50.09	63.11
SAM [34]	✓	✓		55.86	63.33	64.37	61.19	64.49	76.70	52.96	64.72
SAGM [105]	✓	✓		56.81	65.10	65.33	62.08	71.17	72.05	55.38	66.20
SWAD [10]	✓	✓		55.63	63.74	63.55	60.97	70.72	72.94	53.45	65.70
EoA [5]	✓	✓		55.63	64.93	64.68	61.75	69.20	77.27	**58.71**	68.39
RNA-Net [86]	✓	✓		55.37	64.20	62.25	60.61	67.45	68.32	54.78	63.52
SimMMDG [23]	✓	✓		**57.24**	65.07	63.55	61.95	72.75	76.14	54.59	67.83
CMRF (ours)	✓	✓		56.55	**68.13**	**67.04**	**63.91**	**76.45**	**82.39**	56.88	**71.91**
Base	✓		✓	55.86	67.47	59.34	60.89	72.83	77.84	43.58	64.75
SAM [34]	✓		✓	58.85	67.33	63.96	63.38	74.27	78.98	46.79	66.68
SAGM [105]	✓		✓	57.64	66.70	64.67	63.00	76.78	75.10	45.80	65.89
SWAD [10]	✓		✓	59.79	67.33	62.47	63.20	75.82	78.33	51.90	68.68
EoA [5]	✓		✓	62.99	**68.89**	63.76	65.21	74.45	80.68	53.13	69.42
RNA-Net [86]	✓		✓	54.21	64.80	59.31	59.44	74.56	75.39	44.90	64.95
SimMMDG [23]	✓		✓	57.03	66.67	63.86	62.82	77.90	78.98	**57.80**	71.56
CMRF (ours)	✓		✓	**65.28**	67.87	**64.89**	**66.01**	**81.16**	**81.25**	55.50	**72.64**
Base		✓	✓	49.42	55.60	54.41	53.14	52.89	55.11	40.92	49.64
SAM [34]		✓	✓	54.48	59.87	57.90	57.42	54.71	59.66	47.21	53.86
SAGM [105]		✓	✓	55.76	61.32	60.28	59.11	55.90	61.03	47.48	54.80
SWAD [10]		✓	✓	51.32	61.74	61.05	58.04	54.71	59.76	52.00	55.49
EoA [5]		✓	✓	52.41	60.67	61.81	58.30	55.43	58.97	52.29	55.56
RNA-Net [86]		✓	✓	50.89	54.24	55.90	53.68	53.11	59.32	43.82	52.08
SimMMDG [23]		✓	✓	55.86	64.60	59.34	59.93	57.88	60.79	48.62	55.76
CMRF (ours)		✓	✓	**57.24**	**64.94**	**66.12**	**62.76**	**59.06**	**61.79**	**55.04**	**58.49**
Base	✓	✓	✓	54.71	67.20	61.70	61.20	70.29	71.25	53.57	65.07
SAM [34]	✓	✓	✓	56.78	65.20	62.22	61.40	75.36	73.68	57.34	68.79
SAGM [105]	✓	✓	✓	57.76	67.12	61.78	62.22	76.56	75.48	56.92	69.65
SWAD [10]	✓	✓	✓	55.84	68.21	64.90	62.98	75.78	74.95	58.02	69.58
EoA [5]	✓	✓	✓	57.93	68.53	68.78	65.08	76.09	76.95	57.19	70.08
RNA-Net [86]	✓	✓	✓	56.25	63.47	59.72	59.81	71.89	70.88	54.58	65.78
SimMMDG [23]	✓	✓	✓	**62.08**	66.13	64.40	64.20	76.27	77.70	56.42	70.13
CMRF (ours)	✓	✓	✓	61.84	**70.13**	**70.12**	**67.36**	**78.26**	**79.54**	**60.09**	**72.44**

separate domains: H, A, and C. The HAC dataset includes 3381 video clips sourced from the internet, with approximately 1000 samples per domain. Both datasets offer three modalities: video, audio, and optical flow.

Baselines. In our experiments, we compare our CMRF with seven different baselines that can be divided into four groups: 1) Base, naive multi-modal joint training without any domain generalization strategies, 2) SAM [34] and SAGM [105], searching for flat minima in parameter loss landscapes, 3) SWAD [10] and EoA [5], ensemble-based methods for flat minima, and 4) RNA-Net [86] and SimMMDG [23], domain generalization methods specifically designed for MMDG. SAM, SAGM, SWAD and

Table 4.3: Multi-modal **single-source** DG with video, flow and audio three modalities on EPIC-Kitchens and HAC datasets.

		EPIC-Kitchens							**HAC**						
	Source:	D1		D2		D3			H		A		C		
Method	Target:	D2	D3	D1	D3	D1	D2	*Avg*	A	C	H	C	H	A	*Avg*
Base		56.80	53.08	47.36	59.65	55.63	56.93	54.91	64.20	39.45	64.85	52.29	57.97	65.90	57.44
SAM [34]		54.40	55.24	49.65	61.40	54.94	65.33	56.83	67.61	44.04	66.67	**60.09**	60.14	61.36	59.98
SAGM [105]		53.11	57.32	50.46	60.12	56.79	65.10	57.15	67.86	45.31	64.90	57.35	64.10	63.16	60.45
SWAD [10]		57.46	56.92	50.46	63.33	56.25	64.58	58.17	68.43	43.79	68.32	57.35	62.80	67.37	61.34
EoA [5]		58.40	57.39	51.26	64.58	55.17	63.33	58.35	68.18	44.95	69.94	56.88	**67.39**	69.02	62.73
RNA-Net [86]		50.32	51.27	48.90	61.34	53.76	55.89	53.58	62.35	43.24	64.21	53.46	55.37	66.82	57.57
SimMMDG [23]		54.13	**57.90**	50.57	63.04	**60.69**	64.27	58.43	64.77	39.44	71.38	50.46	60.14	70.77	59.49
CMRF (ours)		**60.80**	56.78	**55.17**	**64.99**	57.24	**65.73**	**60.12**	**68.75**	**46.33**	**73.55**	58.26	65.22	**72.46**	**64.09**

EoA are initially designed for uni-modal DG and we extent them into MMDG. For all methods, we follow [127] and select the model with best validation (in-domain) accuracy to evaluate generalization on test (out-of-domain) data. We report the Top-1 accuracy for all results.

Implementation Details. In our framework, we conduct experiments across three modalities: video, audio, and optical flow, adhering to the implementation described in [23]. We leverage the MMAction2 toolkit [18] for our experimental setup. To encode visual information, we utilize the SlowFast network [33], initialized with pre-trained weights on Kinetics-400 [57]. For the audio encoder, we employ ResNet-18 [44], initialized with weights from the VGGSound pre-trained checkpoint [13]. The optical flow encoder uses the SlowFast network's slow-only pathway with Kinetics-400 pre-trained weights. The dimensions of the uni-modal feature $\boldsymbol{h}$ are 2304 for video, 512 for audio, and 2048 for optical flow. For the projector $Proj_k(\cdot)$, we implement a multi-layer perceptron with two hidden layers of size 2048 and output size 128. We use the Adam optimizer [61] with a learning rate of 0.0001 and a batch size of 16. The scalar temperature parameter τ is set to 0.1. Additionally, we set $\lambda_1 = 2.0$, $\lambda_2 = \lambda_3 = 3.0$, α in the Beta distribution to 0.1, and the SMA start iteration t_0 to 400 for EPIC-Kitchens and 100 for HAC respectively. All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU with a 3.9-GHz Intel Core i9-12900K CPU.

Table 4.4: The average results of uni-modal performance comparison under multi-modal multi-source DG on EPIC-Kitchens with different modality combinations.

	Video	Audio	Video-Audio	Video	Flow	Video-Flow	Flow	Audio	Flow-Audio
Uni-video	58.73	-	-	58.73	-	-	-	-	-
Uni-flow	-	-	-	-	58.30	-	58.30	-	-
Uni-audio	-	40.04	-	-	-	-	-	40.04	-
Base	56.65	38.62	59.63	55.28	55.78	60.89	54.86	39.42	53.14
SAM [34]	58.80	37.77	61.19	59.76	56.05	64.05	56.82	40.35	57.42
EoA [5]	57.54	39.70	61.75	57.49	57.17	65.21	57.32	40.14	58.30
SimMMDG [23]	59.43	38.43	61.95	57.02	55.60	62.82	58.21	40.03	59.93
CMRF (ours)	**60.66**	**43.13**	**63.91**	**59.83**	**58.33**	**66.01**	**59.63**	**43.58**	**62.76**

Figure 4.3: Ablations of each module on EPIC-Kitchens with video and audio data. DL: distillation loss, UCL: uni-modal classification loss, CL: contrastive loss, AW: adaptive weight, SMA: simple moving average.

DL	UCL	CL	AW	SMA	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
					54.94	62.26	61.70	59.63
✓					55.63	63.87	62.14	60.55
	✓				53.10	64.12	64.70	60.64
✓	✓				52.75	66.33	65.21	61.43
✓	✓	✓			55.79	65.65	63.92	61.79
✓	✓	✓	✓		53.84	66.79	66.14	62.26
✓	✓	✓		✓	55.79	67.53	65.21	62.84
✓	✓	✓	✓	✓	**56.55**	**68.13**	**67.04**	**63.91**

The model is trained with 15 epochs, taking two hours.

4.6.2 Main Results

Multi-modal multi-source DG. Tab. 4.2 illustrate the results of our CMRF and all baselines on EPIC-Kitchens and HAC under multi-modal multi-source domain generalization setting, where the models are trained on multiple source domains and test on one target domain. We conduct experiments by combining any two modalities, as well as all three modalities, to validate the generalization of our method. As we can see from Tab. 4.2, our CMRF outperforms all baselines on almost all settings and achieves great improvement on the average results (by up to 3.52% with video-

Figure 4.4: Ablation studies on interpolated representations on HAC with video and audio data. SM dis: self-modal distillation, CM dis: cross-modal distillation, Fixed Mix: interpolations with fixed mixing ratio (0.5-0.5).

Method	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
SM dis	74.37	80.68	56.42	70.49
CM dis	75.72	78.85	54.13	69.57
Fixed Mix	75.26	81.81	53.21	70.09
Rand Mix (ours)	**76.45**	**82.39**	**56.88**	**71.91**

audio modalities on HAC). The uni-modal DG methods, especially SAGM and EoA, can improve the generalization of multi-modal network to a certain extent, but their improvements are limited as they do not consider modality competition and inconsistent flatness between modalities. Two MMDG methods RNA-Net and SimMMDG also perform less than satisfactory since they do not fully exploit the generalization capability of each modality.

Multi-modal single-source DG. Our CMRF does not requires domain labels for training, making it feasible to perform multi-modal single-source domain generalization, where models are trained on a single source domain and test on other multiple target domains. The results trained with three modalities are presented in Tab. 4.3. Our CMRF still apparently outperforms all baselines on average accuracy, despite being trained only on single-source domain data. For baselines with domain generalization strategies, they can not improve consistently across datasets, e.g., SimMMDG achieves the second best on EPIC-Kitchens but has limited improvement on HAC, showing their unstable generalization and their limitations in the single-source DG setting.

Uni-modal performance in MMDG. As we discussed in Sec. 4.4, exploiting the generalization capability of each modality simultaneously is the key to improving multi-modal domain generalization performance. Therefore, we evaluate the uni-modal performance from multi-modal trained networks to show the superiority of our method. We freeze the trained uni-modal feature extractor and train a linear classifier to test uni-modal performance. The results of average multi-source accuracy on EPIC-Kitchens are shown in Tab. 4.4. We can see that our CMRF not only improves the multi-modal domain generalization, but also greatly promotes its uni-modal domain generalization, even better than that of uni-modal training (60.66% vs. 58.73% and 43.12% vs. 40.04% for video and audio on EPIC-Kitchens), indicating the effectiveness of CMRF to use cross-modal knowledge to promote the generalization of each modality via mitigating modality competition and flattening representation loss landscape between modalities. In Sec 4.6.4, we show the alleviated competition under in-domain performance and flatter region with perturbations. As for baselines, SAM and SimMMDG only enhance the generalization of better modality and EoA just achieves marginal uni-modal improvement, which means they can not utilize the generalization capability of all modalities comprehensively. Detailed results for each test domain and more results on HAC dataset are shown in Sec. 4.6.4.

4.6.3 Ablation Studies

Ablation on each design. Our CMRF contains five main modules: distillation loss $\mathcal{L}^k_{dis}$, uni-modal classification loss $\mathcal{L}^k_{cls}$, multi-modal supervised contrastive loss $\mathcal{L}_{con}$, adaptive weight, and SMA for teacher model. We conduct extensive ablation experiments to verify the effectiveness of each proposed module on EPIC-Kitchens with video-audio data under multi-source domain generalization setting. The results are illustrated in Tab. 5.4. Only applying distillation loss or uni-modal classification loss improves slightly and their combination leads to noticeable increase, highlighting the importance of flattening representation loss landscape between modalities for domain generalization. However, it does not guarantee steady improvement, e.g., the accuracy decreases from 54.94% to 52.75% in D2, D3 → D1 setting. Multi-modal supervised contrastive loss can enhance the average generalization by a small margin. Adaptive weight and using SMA network as teacher can both improve MMDG by a large margin, suggesting that it is necessary to emphasize the more generalized

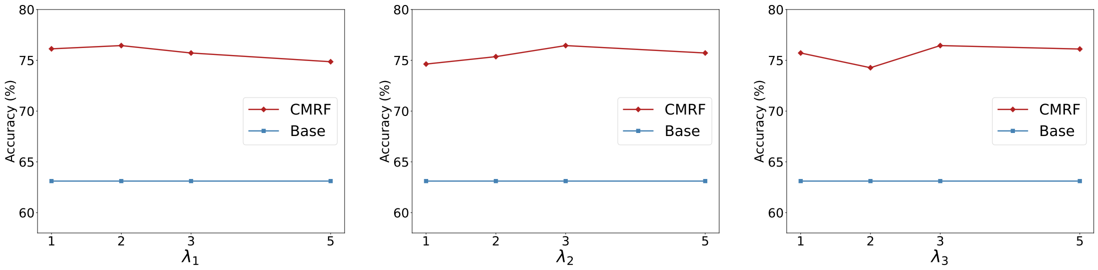


Figure 4.5: Parameter sensitivity analysis on HAC with video and audio data under A, C → H.

modality and obtain more stable distillation signals. Finally, combining all of them achieves the best results for multi-modal domain generalization, hence, each of them is indispensable.

Table 4.5: The average results compared with methods designed for modality competition on HAC with video and audio data under multi-source DG.

	Validation	Test
Base	91.41	63.11
Grad Blending [106]	92.70	66.82
OGM-GE [85]	93.67	64.33
PMR [30]	**94.90**	65.24
CMRF	93.21	**71.91**

Ablation on interpolations. In this section, we mix multi-modal representations in the random ratio generated from Beta distribution as teacher signals, and choose interpolations closer to current modality for distillation, as in Eq. 4.9. We conduct experiments by using different forms of teacher signals to verify our method's effectiveness, as presented in Tab. 4.4. For k-th modality, we set δ to 1, 0, 0.5 for self-modal distillation, cross-modal distillation, and distillation with fixed mixing ratio. Since self-modal distillation can enhance learning for each modality via more generalizable signals, it achieves great performance next to ours. The heterogeneous knowledge between modalities makes cross-mode distillation worse. Fixed mixing ratio only locates one interpolation while our random ratio covers all possible points, resulting in our better performance.

Comparison with methods designed for modality competition. Here, we conduct experiments with three baselines Gradient Blending [106], OGM-GE [85], and PMR [30] for modality competition as we attribute it as one challenge for MMDG. We not only report out-of-domain test accuracy but also in-domain validation results, as shown in Tab. 4.5. We can see that these methods can actually promote their performance on multi-modal validation set since they mitigate the competition. However, they tend to locate at sharp minima and the generalization gap between modalities still makes it hard to build consistent flat minima for different modalities. Hence, their performance increase on test set is limited, while our method achieves significant improvement on both validation and test sets.

Parameter sensitivity. Fig. 4.5 shows the results of different values on loss weights λ_1, λ_2, and λ_3. Since our method uses the moving averaged teacher model for evaluation, it is insensitive to hyperparameters.

4.6.4 More Results

Uni-modal in-domain validation performance. Modal competition refers to the mutual inhibition between modalities in joint training, which is reflected in in-domain performance straightforwardly as studied in previous literature. In Tab.4.6 we give the uni-modal validation results (in-domain) on EPIC-kitchens with video and audio data. Modal competition is manifested in that each single modality of Base performs worse than uni-modal training, which further leads to worse out-of-domain performance as shown in Tab. 4.7. Our method achieves the best uni-modal in-domain performance, indicating that it optimizes modal competition effectively, which in turn improves the generalization ability to other domains as in Tab. 4.4.

Flatness visualization. To evaluate the loss flatness, we can apply low-frequency perturbation from the Gaussian Distribution on representations, where the variance controls the perturbation strength. The magnitude of the performance drop indicates

Table 4.6: Uni-modal validation (in-domain) performance under multi-modal multi-source DG on EPIC-Kitchens dataset with video and audio data.

	Video				Audio			
	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
Uni-modal	79.58	75.58	75.19	76.78	**60.32**	54.29	53.16	55.92
Base	75.78	73.60	72.40	73.93	54.58	52.23	49.11	51.97
SAM	77.03	73.81	73.75	74.86	54.90	51.60	49.67	52.06
EoA	78.94	73.20	75.12	75.75	56.85	52.76	52.45	54.02
SimMMDG	80.86	74.81	74.57	76.75	54.58	53.34	52.90	53.60
CMRF(ours)	**81.26**	**77.21**	**75.69**	**78.05**	58.77	**54.89**	**54.38**	**56.01**

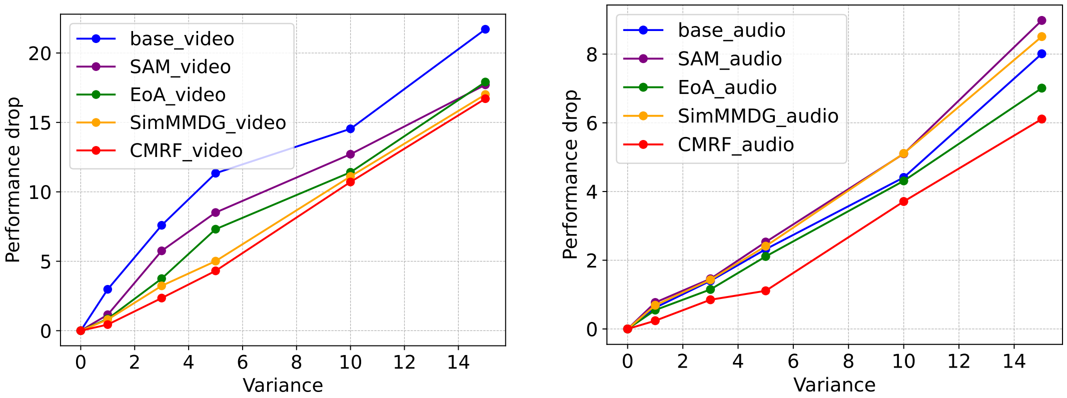


Figure 4.6: Representation space loss flatness evaluation. We apply gaussian noise to the extracted representations to be the domain shifts. The perturbation variance measures the distance between the perturbed representation and the original representation. We use the performance drop against perturbation variance to measure the sharpness of the landscapes around the minimum, where a larger drop indicates a sharp minimum. The experiments are on EPIC-Kitchens with D2, D3 → D1 of video-audio modalities. Left is the performance drop of video while right is the result of audio.

how flat the loss is. The results are shown Figs. 4.6 and 4.7 below. With the increase of Variance, our method has the smallest performance drop on each modality, indicating that our method achieves flatter loss landscape for both modalities simultaneously and in turn provides flatter multi-modal loss landscape.

Uni-modal out-of-domain performance. Here, we give the detailed results of uni-modal performance comparison on EPIC-Kitchens in Tabs. 4.7, 4.8, and 4.9, which form the results in Tab. 4.4 in the main section. The results for HAC dataset are demonstrated in Tabs. 4.10, 4.11, and 4.12. Our method can achieve the best uni-

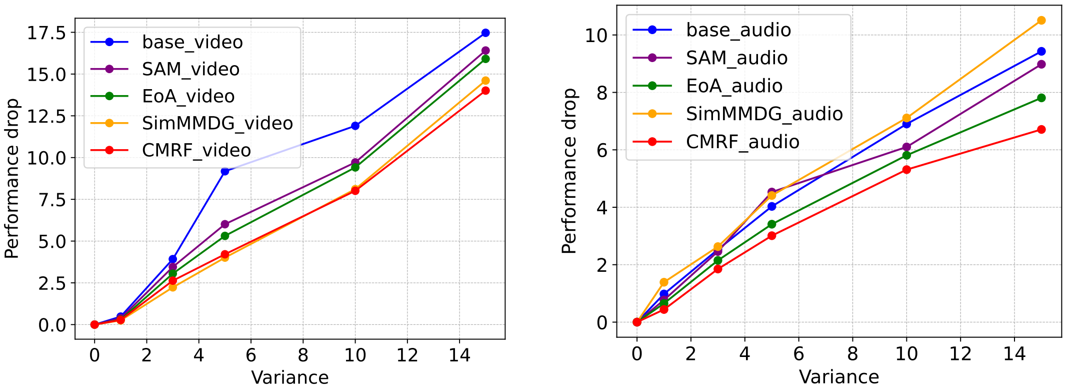


Figure 4.7: Representation space loss flatness evaluation. EPIC-Kitchens with D2, D3 → D1 of flow-audio modalities. Left is the performance drop of flow while right is the result of audio.

Table 4.7: Uni-modal performance under multi-modal multi-source DG on EPIC-Kitchens dataset with video and audio data.

	EPIC-Kitchens							
	Video				Audio			
	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
Uni-video	54.02	**65.60**	56.57	58.73	-	-	-	-
Uni-audio	-	-	-	-	37.01	40.40	42.71	40.04
Base	53.33	62.00	54.62	56.65	36.32	34.60	44.95	38.62
SAM [34]	55.86	61.20	59.34	58.80	33.32	35.87	44.13	37.77
EoA [5]	53.82	63.14	55.67	57.54	**38.16**	37.04	43.55	39.70
SimMMDG [23]	54.67	63.75	59.87	59.43	32.21	34.98	48.12	38.43
CMRF (ours)	**56.79**	64.10	**61.09**	**60.66**	37.94	**43.32**	**48.12**	**43.13**

modal, as well as multi-modal, performance on both datasets with various modality combinations.

Validation and test comparison with uni-modal training. In Tab. 4.13 and Tab. 4.14, we report the in-domain validation and out-of-domain test results on EPIC-kitchens and HAC datasets for each modality. We can see that for each modality, its validation performance is strongly positive correlated to its test performance, i.e., modalities that perform better on the validation set usually perform better on the test set. This provides empirical support for us to use validation set accuracy in Eq. 4.11 to evaluate the generalization ability of different modalities.

Table 4.8: Uni-modal performance under multi-modal multi-source DG on EPIC-Kitchens dataset with video and optical flow data.

	EPIC-Kitchens							
	Video				Flow			
	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
Uni-video	54.02	**65.60**	56.57	58.73	-	-	-	-
Uni-flow	-	-	-	-	**56.55**	62.00	56.36	58.30
Base	47.82	61.47	56.57	55.28	52.18	60.53	54.62	55.78
SAM [34]	54.94	63.87	60.47	59.76	52.64	59.47	56.03	56.05
EoA [5]	51.67	63.33	57.48	57.49	53.04	62.13	56.34	57.17
SimMMDG [23]	50.54	60.76	59.77	57.02	50.33	62.89	53.58	55.60
CMRF (ours)	**55.63**	62.13	**61.74**	**59.83**	53.79	**63.10**	**58.11**	**58.33**

Table 4.9: Uni-modal performance under multi-modal multi-source DG on EPIC-Kitchens dataset with optical flow and audio data.

	EPIC-Kitchens							
	Flow				Audio			
	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
Uni-flow	**56.55**	62.00	56.36	58.30	-	-	-	-
Uni-audio	-	-	-	-	37.01	40.40	42.71	40.04
Base	51.72	57.73	55.13	54.86	36.32	38.00	43.94	39.42
SAM [34]	53.56	60.00	56.90	56.82	37.70	38.93	44.43	40.35
EoA [5]	54.43	59.87	57.67	57.32	38.16	40.40	41.85	40.14
SimMMDG [23]	56.27	61.58	56.79	58.21	35.82	36.49	47.78	40.03
CMRF (ours)	56.27	**63.37**	**59.24**	**59.63**	**40.00**	**41.47**	**49.28**	**43.58**

Table 4.10: Uni-modal performance under multi-modal multi-source DG on HAC dataset with video and audio data.

	HAC							
	Video				Audio			
	A, C → H	H, C → A	H, A → C	*Avg*	A, C → H	H, C → A	H, A → C	*Avg*
Uni-video	73.29	77.11	53.80	68.07	-	-	-	-
Uni-audio	-	-	-	-	28.26	38.09	**32.11**	32.81
Base	72.83	72.72	**57.26**	67.60	**31.16**	36.50	26.06	31.24
SAM [34]	71.84	78.41	55.13	68.46	30.25	39.20	25.23	31.56
CMRF (ours)	**74.64**	**83.52**	53.46	**70.54**	30.43	**44.32**	29.82	**34.86**

Table 4.11: Uni-modal performance under multi-modal multi-source DG on HAC dataset with video and optical flow data.

	HAC							
	Video				Flow			
	A, C → H	H, C → A	H, A → C	*Avg*	A, C → H	H, C → A	H, A → C	*Avg*
Uni-video	73.29	77.11	**53.80**	68.07	-	-	-	-
Uni-flow	-	-	-	-	57.97	58.52	**43.12**	53.20
Base	72.10	74.43	46.33	64.29	56.16	53.98	35.78	48.64
SAM [34]	74.64	78.98	49.08	67.57	53.62	50.00	37.15	46.92
CMRF (ours)	**77.90**	**79.84**	48.33	**68.69**	**63.04**	**62.50**	37.78	**54.44**

Table 4.12: Uni-modal performance under multi-modal multi-source DG on HAC dataset with optical flow and audio data.

	HAC							
	Flow				Audio			
	A, C → H	H, C → A	H, A → C	*Avg*	A, C → H	H, C → A	H, A → C	*Avg*
Uni-flow	57.97	**58.52**	43.12	53.20	-	-	-	-
Uni-audio	-	-	-	-	28.26	38.07	32.11	32.81
Base	55.86	56.82	41.50	51.39	27.35	37.34	26.15	30.28
SAM	60.51	55.13	**48.62**	54.75	29.16	40.04	30.23	32.14
CMRF (ours)	**61.59**	57.95	47.49	**55.68**	**31.88**	**41.48**	**33.03**	**35.46**

Table 4.13: Uni-modal validation performance vs. test performance on EPIC-Kitchens dataset.

	Validation				Test			
	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Avg*
Video	79.58	75.58	75.19	76.78	54.02	65.60	56.57	58.73
Flow	74.94	72.04	72.57	73.18	56.55	62.00	56.36	58.30
Audio	60.32	54.29	53.16	55.92	37.01	40.40	42.71	40.04

Table 4.14: Uni-modal validation performance vs. test performance on HAC dataset.

	Validation				Test			
	A, C → H	H, C → A	H, A → C	*Avg*	A, C → H	H, C → A	H, A → C	*Avg*
Video	90.10	88.66	93.58	90.78	73.29	77.11	53.80	68.07
Flow	74.11	72.87	80.53	78.54	57.97	58.52	43.12	53.20
Audio	56.09	49.19	55.09	53.46	28.26	38.07	32.11	32.81

4.7 Remarks

In this chapter, we analyze the behavior of multi-modal domain generalization and find that modality competition and discrepant uni-modal flatness restrict the generalization capability of multi-modal network. To address these challenges, we propose cross-modal representation flattening (CMRF) to construct consistent flat regions in a shared representation-space loss landscape. Our method builds interpolations by mixing multi-modal representations from moving averaged teacher model and use feature distillation to optimize the high-loss regions between modalities. Our extensive experiments on two benchmark datasets demonstrate the effectiveness of our method to promote multi-modal domain generalization, as well as uni-modal domain generalization in multi-modal network.

Chapter 5

Overcome Modal Bias in Multi-modal Federated Learning via Balanced Modality Selection

Selecting proper clients to participate in each federated learning (FL) round is critical to effectively harness a broad range of distributed data. Existing client selection methods simply consider the mining of distributed uni-modal data, yet, their effectiveness may diminish in multi-modal FL (MFL) as the modality imbalance problem not only impedes the collaborative local training but also leads to a severe global modality-level bias. We empirically reveal that local training with a certain single modality may contribute more to the global model than training with all local modalities. To effectively exploit the distributed multiple modalities, we propose a novel Balanced Modality Selection framework for MFL (BMSFed) to overcome the modal bias. On the one hand, we introduce a modal enhancement loss during local training to alleviate local imbalance based on the aggregated global prototypes. On the other hand, we propose the modality selection aiming to select subsets of local modalities with great diversity and achieving global modal balance simultaneously. Our extensive experiments on audio-visual, colored-gray, and front-back datasets showcase the superiority of BMSFed over baselines and its effectiveness in multi-modal data exploitation.

5.1 Introduction

Federated learning (FL) [75] aims to collaboratively learn data that has been collected by, and resides on, a number of remote devices or servers. FL stands to develop top-performing models by aggregating knowledge from numerous edge clients [20, 104], which relies on the iterative interaction among participating clients and the server. However, comprehensively employing the information from all clients can be exceptionally difficult due to the client heterogeneity (where clients have inherent different data distributions) and resource limitations [26, 89].

Random sampling [102, 68] from available clients has been widely used in FL to satisfy some practical restrictions, e.g., limited communication bandwidth [82, 103] and computing capacities [50]. To improve the information exploitation of all clients, extensive research has been conducted on effective client selection strategies [22, 118]. Despite the success of traditional client selection methods in uni-modal FL, their effectiveness diminishes when dealing with clients with multi-modal data as the inter-modal interactions during the MFL training are neglected. According to [106, 49], there may exist inconsistent learning paces for different modalities in multi-modal joint training, i.e., *modality imbalance*, which not only impedes the collaborative local training but also leads to a severe modal bias for global model in MFL. As illustrated in table 5.1, audio modality significantly outperforms visual modality in CREMA-D and AVE datasets during local training and the aggregated model still suffers from it. However, two well-designed client selection methods (pow-d [16] and DivFL [8]) only obtain severely limited improvement over random sampling (FedAvg) on the multi-modal global model. We can see that client selection methods achieve

Table 5.1: Performance of various client selection methods in MFL under IID setting. A and V denote uni-audio and uni-visual while A-V means the multi-modal result. 'Local' represents that a client is trained based its local data without aggregation. A strong modal bias of global model exists on the two datasets.

Dataset	CREMA-D [9]			AVE [99]		
Method	A	V	A-V	A	V	A-V
Local	41.9	20.4	39.6	33.4	16.7	35.2
FedAvg	51.2	20.6	50.7	61.1	26.8	62.2
pow-d [16]	51.5	20.4	50.5	61.9	26.9	62.5
DivFL [8]	**52.3**	21.1	51.7	**62.7**	25.3	63.3
FedAvg-0.2	50.6	28.6	52.4	60.6	29.6	63.4
FedAvg-0.5	50.5	34.6	55.7	58.7	30.0	60.7
FedAvg-0.8	48.1	**50.9**	61.2	56.4	31.8	58.5
BMSFed	51.0	41.9	**64.5**	59.7	**40.2**	**64.7**

the best uni-audio performance while uni-visual performance even drops sometimes, which means existing client selection scheme heavily relies on the better modality, while ignoring the importance of improving weak modalities that also has potential for global model aggregation. Based on the above analysis, a pivotal question arises: *Can we design a new selection scheme in MFL that can overcome the modal bias and exploit each modality comprehensively?*

To answer this question, we investigate the interactions between different modalities via randomly discarding the data from one modality (audio or visual) on part of clients, which is inspired from modality dropout [3, 114] that drops a specific modal data during training for regularization. The results are shown in rows 7-9 in table 5.1, where '-x' denotes randomly discarding a modality on a client with probability 'x'. We can see that dropping with a certain probability can improve the global multi-modal performance on both datasets (e.g., FedAvg-0.2), and the main reason comes from the dramatic improvement of visual modality. This phenomenon suggests that performing uni-modal training can unleash its potential without being inhibited by another modality, and **uni-modal local training may contribute more to the global model than multi-modal training** on some clients. As the dropping ratio

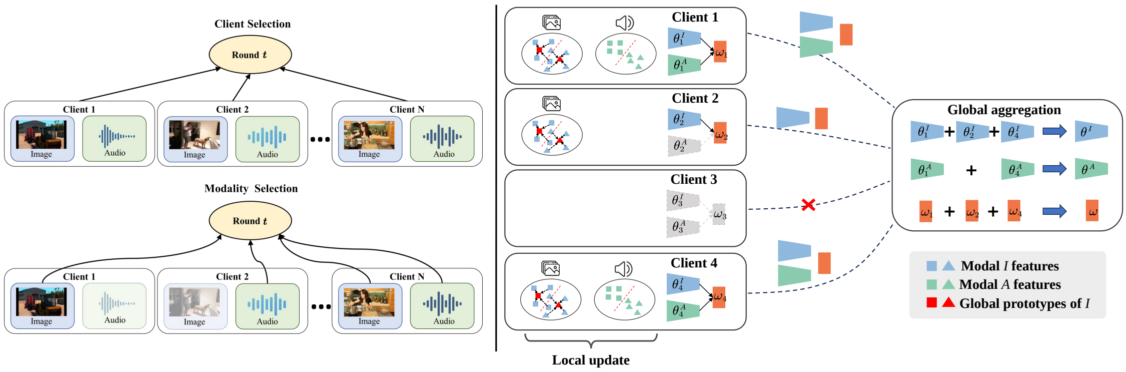


Figure 5.1: **Left:** Traditional client selection in FL aims to sample a client subset in each round while our modality selection considers each local modality as the sampling unit. **Right:** The paradigm of BMSFed with four clients. The global prototypes are used to enhance the weak modality during local update. Only networks corresponding to the selected modalities will be uploaded to the server for aggregation.

increases, although the visual modality still improves, multi-modal performance may decline as the audio modality declines as shown in columns 5 and 7, suggesting that we should carefully control *which modalities on each client should be involved in training and aggregation to make the contribution most* to the global model.

According to the above investigations, we propose a novel Balanced Modality Selection scheme for MFL (BMSFed) to mitigating the modal bias and comprehensively exploit the diverse information from all modalities. Specifically, instead of selecting a subset of clients, we treat each modality on the local side as a selection unit, as demonstrated in fig. 5.1. Our BMSFed mainly contains two parts: Firstly, we intend to alleviate the local modality imbalance by introducing a modal enhancement loss based on aggregated global prototypes to promote the performance of weak modality. Secondly, we complete the modality selection by building two separated submodular functions for selecting multi-modal clients (training with multi-modal data) and uni-modal clients (training with selected uni-modal data) respectively. Inspired from [8], the criterion is to select modalities that are most representative on the gradients while also alleviate the global modal bias. A simple yet effective conflict resolution strategy is devised to ensure the validity of modality selection and keep modal balance on global model simultaneously.

The main contributions of the section are summarized as follows:

1. We empirically analyze the modality imbalance problem in MFL and reveal that uni-modal training on some clients may contribute more to the global model than multi-modal training.

2. Based on the analysis, we propose a novel Balanced Modality Selection scheme for MFL (BMSFed) to comprehensively exploit all modalities via a modal enhancement loss and representative modality selection to overcome the global modal bias.

3. We conduct comprehensive experiments on audio-visual, colored-gray, and front-back datasets, and considering the statistical heterogeneity and modality incongruity problems in MFL, to validate the superiority of our BMSFed.

5.2 Background and Related Works

5.2.1 Multi-modal Federated Learning

In MFL, each client has one or multiple modalities of data. Without loss of generality, we consider two input modalities, which are denoted by A and I respectively in MFL. There are a set V of N clients, $V = [N]$, respectively owning datasets $\mathcal{D}_i = \left\{\boldsymbol{X}_i^A, \boldsymbol{X}_i^I, \boldsymbol{y}\right\}$, $i \in [N]$. A typical federated learning objective is the average of each client's local loss function:

$$\min f\left(\theta\right) = \sum_{k=1}^{N} p_k F_k\left(\theta\right) \tag{5.1}$$

where $\theta = \left\{\theta^A, \theta^I, \omega\right\}$ denotes the model parameters. θ^A and θ^I represent the encoder parameters of modality A and I. ω is the parameter of fusion classifier. p_k is a predefined weight. F_k is each client's local loss (cross entropy (CE) loss for classification task in this section).

Statistical heterogeneity [55, 66] is a widely concerned challenge in uni-modal FL. To tackle this issue, FedProx [67] uses a proximal term to stabilize model aggregation. FedProto [98] shares class prototypes to regularize the learning of local models. In MFL, modality incongruity [134, 128] (clients consist of different modalities combinations), as well as statistical heterogeneity [116], are all considered. Yu et al. [128] propose CreamFL to align the representations between different clients and different modalities via communicating knowledge on a public dataset. Chen et al. [14] introduce FedMSplit to split local models into several components and aggregate them by their correlations. However, they still focus on the heterogeneity, but ignore the interaction between private data of different modalities, which limits their information exploitation.

5.2.2 Client Selection and Submodular Function

Client selection [22, 118] is a critical issue for FL especially when the communication cost with all devices is prohibitively high, which has been extensively studied in uni-modal FL. Cho et al. [16] propose Power-of-Choice to select clients with largets local loss. Balakrishnan et al. [8] propose to select a subset of clients with great diversity.

Diverse client selection via submodularity. Maximizing a submodular function is reported to improve the diversity and reduce the redundancy of a subset. This property makes it appropriate for client selection in FL. If a function F is submodular, it should satisfy: given a finite ground set V of size N, $F\left(A \cup \{v\}\right) - F\left(A\right) \geqslant F\left(B \cup \{v\}\right) - F\left(B\right)$, for any $A \subseteq B \subseteq V$ and $v \in V \backslash B$. The marginal utility of an element v w.r.t. a subset A is denoted as $F\left(v|A\right) = F\left(A \cup \{v\}\right) - F\left(A\right)$, which can represent the importance of v to A. The client selection via submodular maximization can be expressed following [8]: find a subset S of clients whose aggregated gradients can approximate the full aggregation from all clients:

$$\sum_{k\in[N]} \nabla F_k\left(v^k\right) = \sum_{k\in[N]} \left[\nabla F_k\left(v^k\right) - \nabla F_{\sigma(k)}\left(v^{\sigma(k)}\right)\right] + \sum_{k\in S} \gamma_k \nabla F_k\left(v^k\right) \tag{5.2}$$

where σ maps $V \to S$ and the gradient $\nabla F_k\left(v^k\right)$ from client k is approximated by the gradient from a selected client $\sigma(k) \in S$. For $i \in S$, let $C_i \triangleq \{k \in V | \sigma(k) = i\}$, and therefore $\gamma_i \triangleq |C_i|$. Take the norms and apply triangular inequality after subtracting the second term from both sides, we can obtain a relaxed objective $G(S)$ for minimizing the approximation error:

$$\left\| \sum_{k\in[N]} \nabla F_k\left(v^k\right) - \sum_{k\in S} \gamma_k \nabla F_k\left(v^k\right) \right\| \leqslant \sum_{k\in[N]} \min_{i\in S} \left\|\nabla F_k\left(v^k\right) - \nabla F_i\left(v^i\right)\right\| \triangleq G(S) \tag{5.3}$$

Minimizing $G(S)$ can be seen as maximizing the well-known submodular function, i.e., the facility location function [19]. The submodular maximizing problem is NP-hard but can be approximated via the greedy [79] or stochastic greedy algorithm [76]:

$$S \leftarrow S \cup k^*, k^* \in \underset{k\in\mathrm{rand}(V\backslash S,\mathrm{size}=s)}{\arg\max} \left[\bar{G}(S) - \bar{G}(\{k\} \cup S)\right] \tag{5.4}$$

$\bar{G}$ represents a constant minus the negation of G.

Although these methods make great improvement in uni-modal FL, the selection strategy is under-explored in MFL and we reveal that traditional client selection approaches cannot address the severe modal bias in MFL.

5.2.3 Imbalanced Multi-modal Learning

Modality imbalance indicates the inconsistent learning progress of different modalities in multi-modal learning [106, 49]. Peng et al. [85] propose OGM-GE to alleviate the inhibitory effect on weak modality by slowing down the dominant modality. Fan et al. [30] further build a non-parametric classifier by class centroids to adjust the update direction of weak modality. In this section, we aim to power each modality of all clients by a meticulously designed modality selection strategy in each round of training.

5.3 Methodology

In this section, we introduce BMSFed that contains the local imbalance alleviation and balanced modality selection.

5.3.1 Local Imbalance Alleviation

As discussed in Table 5.1, the multi-modal training on each client may suffer from severe modality imbalance, leading to inadequate information exploitation on the local side and consequently incurring the modal bias on the global model. Therefore, we first try to alleviate the imbalance during local training.

Inspired from [30], we can facilitate learning of weak modality by modulating the gradient direction and magnitude. For the i-th client with the data $\mathcal{D}_i = \left\{\boldsymbol{X}_i^A, \boldsymbol{X}_i^I, \boldsymbol{y}\right\}$, the local prototype for class j of each modality is defined as the mean value of representations:

$$c_{i,j}^I = \frac{1}{|\mathcal{D}_{i,j}|} \sum_{x^I \in \mathcal{D}_{i,j}} h_i\left(\theta_i^I; x^I\right), c_{i,j}^A = \frac{1}{|\mathcal{D}_{i,j}|} \sum_{x^A \in \mathcal{D}_{i,j}} h_i\left(\theta_i^A; x^A\right) \tag{5.5}$$

where $\mathcal{D}_{i,j}$ denotes the samples belonging to j-th class in client i. h_i is the function of encoder. Considering the heterogeneity across clients, we aggregate the local prototypes to a global prototype as:

$$c_j^{GI} = \frac{1}{|\mathcal{N}_j|} \sum_{i \in \mathcal{N}_j} \frac{|\mathcal{D}_{i,j}|}{N_j} c_{i,j}^I, c_j^{GA} = \frac{1}{|\mathcal{N}_j|} \sum_{i \in \mathcal{N}_j} \frac{|\mathcal{D}_{i,j}|}{N_j} c_{i,j}^A \tag{5.6}$$

where $\mathcal{N}_j$ denotes the set of clients that have class j and N_j is the number of instances belonging to class j over all clients. And then, we introduce the modal enhancement loss (ME) for client k based on the global prototype:

$$\begin{aligned}\mathcal{L}_{ME}^{k}\left(v_I^k\right) &= -\mathbb{E}_{\left(x_i^I,y\right)\in\mathcal{D}_k}\log\left[\frac{\exp\left(-d\left(z_i^I,c_y^{GI}\right)\right)}{\sum_{j=1}^{Y}\exp\left(-d\left(z_i^I,c_j^{GI}\right)\right)}\right]\\ \mathcal{L}_{ME}^{k}\left(v_A^k\right) &= -\mathbb{E}_{\left(x_i^A,y\right)\in\mathcal{D}_k}\log\left[\frac{\exp\left(-d\left(z_i^A,c_y^{GA}\right)\right)}{\sum_{j=1}^{Y}\exp\left(-d\left(z_i^A,c_j^{GA}\right)\right)}\right]\end{aligned} \tag{5.7}$$

where $d\left(\cdot,\cdot\right)$ is the distance function (Euclidean distance), z_i^I is the representation of x_i^I, i.e., $z_i^I = h_i\left(\theta_i^I; x_i^I\right)$. Y is the class number. v_A^k and v_I^k indicate the corresponding modal data in client k. Hence, the local loss should be (data superscripts are omitted for simplicity):

$$F_k\left(v_A,v_I\right) = \begin{cases}\mathcal{L}_{CE}^{k}\left(v_A,v_I\right)+\gamma^k\mathcal{L}_{ME}^{k}\left(v_A\right), & \rho_I^k \leqslant 1\\ \mathcal{L}_{CE}^{k}\left(v_A,v_I\right)+\beta^k\mathcal{L}_{ME}^{k}\left(v_I\right), & \rho_I^k > 1\end{cases} \tag{5.8}$$

where ρ_I^k indicates the local imbalance ratio for client k calculated based on the modality-wise ground truth prediction, and $\rho_I^k > 1$ means modality A outperforms modality I and vice verse. $\gamma^k, \beta^k \in (0,1)$ are the modulation coefficients. The calculation details for γ^k, β^k and ρ_I^k are given below.

Prototype. The prototype is the centroid of the representations for each class. Therefore, the local prototype for class j is calculated as (for modal I):

$$c_j^I = \frac{1}{N_j}\sum\nolimits_{i=1}^{N_j} z_{j_i}^I \tag{5.9}$$

where N_j is the number of samples with class j in local side.

When the sever receives the local prototypes from each client, the global prototypes are aggregated according to the sample numbers from each client:

$$c_j^{GI} = \frac{1}{\sum_{k=1}^{N} N_j|_k} \sum_{k=1}^{N} c_j^I|_k \cdot N_j|_k \tag{5.10}$$

where $N_j|_k$ is the number of samples with class j in client k, $c_j^I|_k$ is the prototype of class j in client k.

Imbalance ratio ρ_I and coef β^k. In this section, we need the coefficients β^k and γ^k in eq. (5.8) to adjust the strength of local enhancement and imbalance ratio to determine which modality is weak and modulate the modality selection process as in eq. (5.19). Local imbalance ratio of client k is the quotient of average ground-truth logits from two modalities:

$$\begin{aligned} s_i^A &= \sum_{c=1}^{C} 1_{c=y_i} \cdot \mathrm{soft}\max\left(\hat{y}_i^A\right)_c \\ s_i^I &= \sum_{c=1}^{C} 1_{c=y_i} \cdot \mathrm{soft}\max\left(\hat{y}_i^I\right)_c \end{aligned} \tag{5.11}$$

$$\rho_I^k = \frac{\sum_{i \in B_t} s_i^A}{\sum_{i \in B_t} s_i^I} \tag{5.12}$$

where $\hat{y}_i^A, \hat{y}_i^I$ are the logit outputs based on the distance differences from local prototypes. B_t is a random mini-batch at time step t. Then the global imbalance ratio is calculated as:

$$\rho_I = \frac{1}{\sum_{k=1}^{N} n_k} \sum_{k=1}^{N} \rho_I^k \cdot n_k \tag{5.13}$$

According to [30], we design β^k as:

$$\begin{cases} \gamma^k = clip\left(0, \frac{1}{\rho_I^k} - 1, 1\right) & \rho_I^k < 1 \\ \beta^k = clip\left(0, \rho_I^k - 1, 1\right) & \rho_I^k \geqslant 1 \end{cases} \tag{5.14}$$

Different from [30], we apply the global prototype instead of local prototypes to aggregate the knowledge from different clients, and the ME loss is not only applied to multi-modal clients, but also to uni-modal clients (See details in subsection 5.3.2).

5.3.2 Balanced Modality Selection

Considering that different types of modal combinations may be selected to participate in local training in our scheme, we define the participated clients as multi-modal clients and uni-modal clients (e.g., client 2 in Figure 5.1 Right is a uni-modal client with modal I for training while client 1 and 4 are multi-modal clients with both modalities for training). In subsection 5.3.1, we propose the ME loss to facilitate the learning of weak modality, which is originally designed for multi-modal clients. Here, we also apply it on uni-weak-modal clients to realize gradient consistency on different types of clients.

Assume modality I is weak here. Hence, the local loss for multi-modal and uni-modal clients should be:

$$\begin{aligned} \text{multi-modal}: F_k\left(v_A, v_I\right) &= \mathcal{L}_{CE}^k\left(v_A, v_I\right) + \beta^k \mathcal{L}_{ME}^k\left(v_I\right) \\ \text{uni-modal}: F_k\left(v_A\right) &= \mathcal{L}_{CE}^k\left(v_A\right), F_k\left(v_I\right) = \mathcal{L}_{CE}^k\left(v_I\right) + \beta^k \mathcal{L}_{ME}^k\left(v_I\right) \end{aligned} \tag{5.15}$$

Next, following Eq. 5.2, we define the paradigm of modality selection for MFL, aiming to approximate the full gradient aggregation from all clients by the gradient from the customized multi-modal and uni-modal clients:

$$\begin{aligned} \sum_{k\in[N]} \nabla F_k\left(v_A, v_I\right) = \sum_{k\in[N]} &\left[\begin{array}{c} \nabla F_k\left(v_A, v_I\right) - \nabla F_{\sigma_M(k)}\left(v_A, v_I\right) \\ -\nabla F_{\sigma_A(k)}\left(v_A\right) - \nabla F_{\sigma_I(k)}\left(v_I\right) \end{array}\right] \\ + \sum_{k\in S_M} \gamma_k^M \nabla F_k\left(v_A, v_I\right) + \sum_{k\in S_A} &\gamma_k^A \nabla F_k\left(v_A\right) + \sum_{k\in S_I} \gamma_k^I \nabla F_k\left(v_I\right) \end{aligned} \tag{5.16}$$

where σ_M, σ_A and σ_I map $V \rightarrow S_M, S_A, S_I$, the client sets who use multi-modal, uni-A, and uni-I data for training respectively, and $S_M \cap S_A = S_A \cap S_I = S_M \cap S_I = \oslash$. γ_k^M, γ_k^A and γ_k^I have the similar meaning as γ_k in Eq. 5.2.

Since modality I is weak here, we omit the uni-A clients as the multi-modal gradient is dominated by modality A [85], which means we do not need to select uni-A clients for its enhancement. Then, bring Eq. 5.15 to Eq. 5.16 and follow the operations from

Eq. 5.2 to Eq. 5.3, we can obtain:

$$
\begin{aligned}
&\sum_{k\in[N]} \min_{i\in S_M, j\in S_I} \left\| \nabla F_k\left(v_A, v_I\right) - \gamma_i^M \nabla F_i\left(v_A, v_I\right) - \gamma_j^I \nabla F_j\left(v_I\right) \right\| \\
&= \sum_{k\in[N]} \min_{i\in S_M, j\in S_I} \left\| \begin{array}{c} \nabla \mathcal{L}_{CE}^k\left(v_A, v_I\right) + \nabla \beta^k \mathcal{L}_{ME}^k\left(v_I\right) - \nabla \mathcal{L}_{CE}^i\left(v_A, v_I\right) \\ -\nabla \beta^i \mathcal{L}_{ME}^i\left(v_I\right) - \nabla \mathcal{L}_{CE}^j\left(v_I\right) - \nabla \beta^j \mathcal{L}_{ME}^j\left(v_I\right) \end{array} \right\| \\
&\leqslant \sum_{k\in[N]} \min_{i\in S_M} \left\| \nabla \mathcal{L}_{CE}^k\left(v_A, v_I\right) - \nabla \mathcal{L}_{CE}^i\left(v_A, v_I\right) \right\| \\
&+ \sum_{k\in[N]} \min_{i\in S_M, j\in S_I} \left\| \begin{array}{c} \nabla \beta^k \mathcal{L}_{ME}^k\left(v_I\right) - \nabla \beta^i \mathcal{L}_{ME}^i\left(v_I\right) \\ -\nabla \mathcal{L}_{CE}^j\left(v_I\right) - \nabla \beta^j \mathcal{L}_{ME}^j\left(v_I\right) \end{array} \right\| \\
&\triangleq G\left(S_M\right) + G\left(S_M \cup S_I\right)
\end{aligned}
\tag{5.17}
$$

The right-hand side of the first equation is the modality selection formula that aims to select a group of multi-modal clients S_M and uni-modal clients S_I to approximate the aggregated gradients from all clients. However, the joint selection for S_M and S_I is a complex joint optimization problem. Therefore, we decouple this objective into two submodular functions $G\left(S_M\right)$ and $G\left(S_M \cup S_I\right)$ according to triangle inequality, while the full gradient approximation is divided into two parts: the first part uses selected multi-modal CE gradient to fit fully multi-modal CE gradient aggregation, and the second part approximates the fully multi-modal ME gradient aggregation via selected uni-modal CE gradient and both selected multi- and uni-modal ME gradient. The modality-level gradient decoupling converts the complex joint selection to two simply separated selection problems.

Although we can solve the two submodular functions with the stochastic greedy algorithm [76], there are still two issues: (1) the selected client according to $G\left(S_M \cup S_I\right)$ should be specified whether it is uni-modal client or multi-modal client; (2) the separated selection strategy pays less attention to the global modal bias. To address the two problems, we perform the stochastic greedy algorithm for two submodular functions in parallel and propose a simple yet effective conflict resolution strategy to

Algorithm 1: BMSFed.

```
Input: Input data $\mathcal{D}_i = \left\{ \boldsymbol{X}_i^A, \boldsymbol{X}_i^I, y \right\}$, $i \in [N]$, initial model $\theta$,
       hyper-parameters $\chi$, global communication epochs $E$, $e = 1$.
while $e < E$ do
    if $e = 1$ then
        Send $\theta$ to all clients;
        Perform one-step local update for gradients, prototypes and $\rho_I^k$;
        Aggregate global prototypes and $\rho_I$;
    else
        Aggregate global model $\theta$, prototypes $c^G$ and $\rho_I$;
    Select multi-modality for $S_M$ and uni-modality for $S_I$ (or $S_A$) using Eqs. 5.18
      and 5.19;
    Send $\theta$, $c^G$ and $\rho_I$ to selected clients;
    foreach client in selected clients in parallel do
        Perform multi-modal learning in $S_M$ and uni-modal learning in $S_I$ (or $S_A$)
          by with Eqs. 5.8 and 5.15;
        Send gradients, local prototypes and $\rho_I^k$ to server;
```

ensure $S_M \cap S_I = \oslash$ as well as, more importantly, balance the learning of different modalities on global model:

$$S_M \leftarrow S_M \cup k_1^*, k_1^* \in \underset{k \in \mathrm{rand}(V \backslash S_M \backslash S_I, \mathrm{s})}{\arg\max} \left[\bar{G}\left(S_M\right) - \bar{G}\left(\{k\} \cup S_M\right) \right] \tag{5.18}$$

$$\begin{cases} if\ k_1^* = k_2^*, S_M \cup k_2^*; \\ if\ k_1^* \neq k_2^*, \begin{cases} S_I \cup k_2^*, if\ \rho_I^k > \chi \\ S_M \cup k_2^*, if\ \rho_I^k \leqslant \chi \end{cases} \end{cases} \tag{5.19}$$

$$k_2^* \in \underset{k \in \mathrm{rand}(V \backslash S_M \backslash S_I, \mathrm{s})}{\arg\max} \left[\bar{G}\left(S_M \cup S_I\right) - \bar{G}\left(\{k\} \cup S_M \cup S_I\right) \right]$$

For every selection, we randomly sample a subset of clients s. A multi-modal clients k_1^* is selected from s according to $G\left(S_M\right)$ while k_2^* is also selected from s according to $G\left(S_M \cup S_I\right)$. $k_1^* = k_2^*$ means using multi-modal data from this client can contribute most to the global model. When $k_1^* \neq k_2^*$, we allocate it to uni-modal or multi-modal client according to its local imbalance ratio: if it is severely imbalanced, we use its uni-weak-modal data for training and aggregation to alleviate the global modal bias, otherwise we believe that training with its multi-modal data contributes more than uni-modal data. χ is a hyper-parameter.

Discussion. (1) Overcoming the modal bias in our method are twofold: the ME loss alleviates imbalance at local side and the selected uni-modal clients further promote balanced learning of global model. Meanwhile, the diversity coming from two submodular functions ensures the representative information for the global model. (2) We assume I is the weak modality above while in practice, we can determine the weak modality before modality selection via the aggregated global imbalance ratio $\rho_I = \frac{1}{\sum_{k=1}^{N} n_k} \sum_{k=1}^{N} \rho_I^k \cdot n_k$. Overall, the pseudo-code of BMSFed is provided in Algorithm 1. (3) Only the gradients, prototypes and ρ_I^k participate in communication, so there is no privacy issue and similar communication overheads as in traditional FL.

5.4 Evaluation

5.4.1 Datasets and Baselines

Datasets. We conduct experiments on four datasets: (1) **CREMA-D** [9] is an audio-visual dataset for emotion recognition task with total six categories for emotional states. (2) **AVE** [99] is an audio-visual dataset for event localization with 28 event classes, and here we use it to construct a labeled multi-modal classification dataset following [30]. (3) **Colored-and-gray MNIST** (CG-MNIST) [59] is a synthetic dataset based on MNIST [62] with gray-scale and monochromatic images as two modalities, following [112]. (4) **ModelNet40** is one of the Princeton ModelNet datasets [113] with 3D objects of 40 categories. The front and back [96] views are considered as two modalities, following [14].

Baselines. We choose eight baselines for comparison from four categories: (1) three uni-modal FL methods designed for statistical heterogeneity are extended to multi-modal scenarios: FedAvg [75], FedProx [67] and FedProto [98]. (2) Integrating OGM-GE [85] and PMR [30], the solutions for modality imbalance, with FedAvg forms two MFL methods: FedOGM and FedPMR. (3) Two client selection method: Power-of-choice (pow-d) [16] and DivFL [8], evolved from its uni-modal version directly. (4) One MFL method, FedMSplit [14], especially designed for modality incongruity. Compared with these baselines, we demonstrate that an elaborate modality selection strategy is essential to realize comprehensive information exploitation in MFL.

5.4.2 Experimental Settings

For CREMA-D, AVE and ModelNet40, we use ResNet18 [44] as the backbone for audio, visual and flow modalities. Audio data is converted to a spectrogram of size 257x299 for CREMA-D and 257×1,004 for AVE. We randomly choose 3 frames and 4 frames to build image training sets for CREMA-D and AVE respectively. For CG-MNIST, we build a neural network with 4 convolution layers and 1 average pool layer as the encoder, following the setting as in [30]. We choose the simple yet effective fusion method, concatenation [84], to build fusion classifier for all the datasets. We set 20 clients for CREMA-D, AVE and ModelNet40 while the number for CG-MNIST is 30. 5 clients are selected in each communication round for CREMA-D, AVE, ModelNet40 and 6 for CG-MNIST. For IID setting, training data is uniformly distributed to all clients. For non-IID scenarios, we use Dirichlet distribution [46] $Dir(\alpha)$ to split data ($\alpha = 3$ for CREMA-D, AVE, ModelNet40, $\alpha = 2$ for CG-MNIST). The optimizer is SGD [90] for all datasets. Learning rate is initialized at 1e-3 or 1e-2 for CEAMA-D, AVE and ModelNet40 or CG-MNIST and becomes 1e-4 or 1e-3 in the later training stage. The hyper-parameter χ is set to 1.2-2.5 according to datasets and settings. To complete stochastic greedy algorithm for Eqs. 5.18 and 5.19, we use the gradients from the selected clients at current round to update part of the simi-

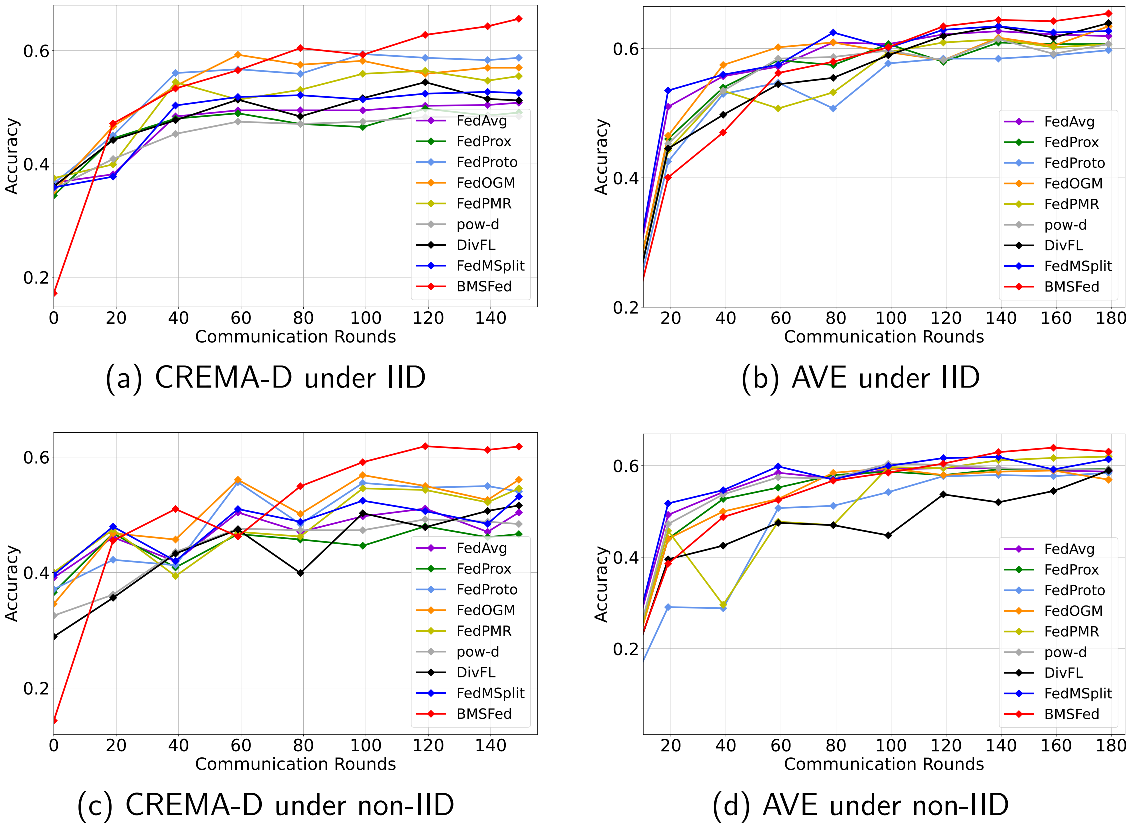
Table 5.2: Comparison results on four datasets. The metric is the top-1 accuracy (%). The best is in **bold**, and the second best is underlined.

Dataset	CREMA-D		AVE		CG-MNIST		ModelNet40	
Method	IID	non-IID	IID	non-IID	IID	non-IID	IID	non-IID
FedAvg	50.7	49.8	62.2	59.7	42.3	41.7	87.2	86.5
FedProx	51.0	49.0	62.6	59.9	42.9	43.6	86.9	87.1
FedProto	58.7	54.0	61.7	58.8	51.5	51.4	87.5	87.2
FedOGM	56.9	56.4	62.8	59.3	57.2	53.0	87.6	87.0
FedPMR	55.5	55.1	63.1	61.6	66.1	63.3	87.6	**87.7**
pow-d	50.5	50.7	62.5	60.0	41.2	40.3	86.8	86.2
DivFL	51.7	50.8	63.3	59.6	43.0	42.1	86.5	86.4
FedMSplit	52.4	51.6	62.4	60.8	43.5	50.9	87.5	87.4
BMSFed	**64.5**	**61.6**	**64.7**	**62.1**	**70.2**	**66.7**	**88.7**	87.5

larity matrix, which is named "no-overheads" in [8]. Except for pow-d, DivFL and BMSFed, other baselines select clients randomly. Each client has two modal data by default. We do all experiments on a workstation with an RTX 3090 GPU, a 3.9-GHZ Intel Core i9-12900K CPU and 64GB of RAM.

5.4.3 Comparison with Baselines

BMSFed effectively improves the performance. As demonstrated in Table 5.2, our BMSFed achieves the best results on the four datasets under both IID and non-IID settings (by up to 5.8% on CREMA-D). Client sampling here (pow-d and DivFL) cannot fully exploit information for all modalities, making its improvement limited or even worse than FedAvg in CG-MNIST. Although FedOGM and FedPMR accomplish modest improvement because of their ability to alleviate modality imbalance, they are not as good as BMSFed since they do not consider the overall performance of the global model. Traditional uni-modal FL methods for statistical heterogeneity (e.g. FedProx) and MFL method for modality incongruity (FedMSplit) only obtain slight improvement. We also illustrate the trend of test accuracy versus the number of



(a) CREMA-D under IID

(b) AVE under IID

(c) CREMA-D under non-IID

(d) AVE under non-IID

Figure 5.2: Test accuracy of BMSFed compared with other baselines on CREMA-D and AVE.

communication rounds on CREMA-D and AVE in Figure 5.2. BMSFed can realize comparable or even faster convergence speeds in CREMA-D and AVE.

BMSFed exploits all modalities comprehensively. To show the effect of our method on addressing modal bias, we report the performance of each modality on CREMA-D and AVE as shown in Table 5.3. The uni-modal performance evaluation follows [30]: a sample is classified into the class corresponding to its nearest prototype. It is clear that BMSFed could considerably improve the performance of weak modality (visual) and mitigate the modality-level bias. Besides, compared with randomly modality abandoning, which significantly reduces audio performance as illustrated in Table 5.1, BMSFed achieves comparable audio performance with other baselines. Although FedProto, FedOGM and FedPMR also alleviate the imbalance, they mainly

Table 5.3: The uni-modal performance comparison on CREMA-D and AVE.

Dataset	CREMA-D				AVE			
Setting	IID		non-IID		IID		non-IID	
Method	A	V	A	V	A	V	A	V
FedAvg	51.2	20.6	50.7	20.2	61.1	26.8	61.4	26.4
FedProx	51.3	20.2	50.1	22.0	60.4	27.1	61.2	26.9
FedProto	50.2	35.3	48.6	39.1	55.7	36.8	59.7	32.8
FedOGM	50.5	35.7	48.8	30.2	58.7	28.8	59.4	29.4
FedPMR	51.5	38.7	50.1	35.9	61.7	39.6	61.7	35.3
pow-d	51.5	20.4	51.6	18.8	61.9	26.9	60.1	27.1
DivFL	**52.3**	21.1	**52.1**	22.7	**62.7**	25.3	61.6	26.3
FedMSplit	52.0	21.8	50.8	21.6	61.3	26.9	**62.3**	28.7
BMSFed	51.0	**41.9**	49.3	**41.4**	59.7	**40.2**	60.2	**38.6**

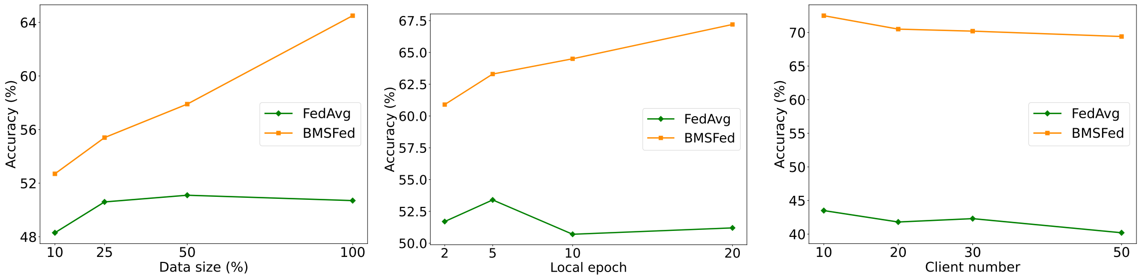
focus on local optimization, resulting in the performance gap between them and our BMSFed on the aggregated model, which further indicates that in MFL, it is important to take both local optimization for each modality and the overall performance for global model into consideration simultaneously.

5.4.4 Ablation Study

Effectiveness of each component. Table 5.4 studies the effect of each BMSFed component. Applying ME loss Eq. 5.7 on random sampling (FedAvg+ME) and the well-designed client selection (DivFL+ME) surpasses their vanilla strategies (FedAvg, DivFL) by a large margin, demonstrating its effectiveness on local enhancement. Comparing BMSFed (64.5% on IID CREMA-D) with 'DivFL+ME' (57.1% on the same setting) also denotes the necessity of balancing different modalities considering the global model via modality selection. To show the importance of aligning feature spaces of weak modality, we replace the global prototypes with local prototypes (BMSFed-local). Global alignment achieves notable improvement (by up to 2% on non-IID AVE). The performance improvement compared with 'FedAvg-0.2,0.5,0.8' exhibits that randomly sampling modalities does not always lead to improvement and

Table 5.4: Ablation study. 'BMSFed-local' uses local prototypes rather than global prototypes for ME loss.

Dataset	CREMA-D		AVE	
setting	IID	non-IID	IID	non-IID
FedAvg	50.7	49.8	62.2	59.7
DivFL	51.7	50.8	63.3	59.6
FedAvg-0.2	52.4	50.1	63.4	61.1
FedAvg-0.5	55.7	55.1	60.7	59.4
FedAvg-0.8	61.2	58.1	58.5	58.7
FedAvg+ME	55.8	54.5	62.8	60.7
DivFL+ME	57.1	55.6	63.0	61.1
BMSFed-local	63.7	60.3	63.4	60.1
BMSFed	**64.5**	**61.6**	**64.7**	**62.1**



(a) CREMA-D (b) CREMA-D (c) CG-MNIST

Figure 5.3: Robustness validation on (a) data size, (b) local epoch and (c) client number under IID setting.

further demonstrates the need of meticulously selecting modalities for information exploitation.

Robustness test. To verify the robustness of our method, we vary three key hyperparameters to build various scenarios: (1) change the data size $|\mathcal{D}_i|$ to allow each client to hold a small amount of data, (2) set different local training epochs, and (3) vary the total client number N. As illustrated in Figures 5.3a, b, our BMSFed consistently outperforms baseline (FedAvg) under various scenarios. More data as well as more local training epochs can bring further improvements to our method, indicating that exploiting the weak modality need more training efforts. Based on the results in Figure 5.3c, our approach can also be generalised to clients with larger

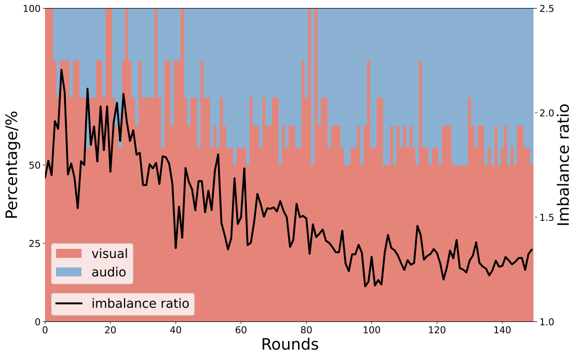


Figure 5.4: Proportional change of audio and visual respectively and the curve of global imbalance ratio during training on CREMA-D under IID setting.

Table 5.5: Performance on CREMA-D and AVE with modality incongruity. 50% of clients have all modal data and 50% of clients only retain data with a single modality (audio or visual).

Dataset	CREMA-D		AVE	
setting	IID	non-IID	IID	non-IID
FedAvg	55.7	55.1	60.7	58.7
FedProx	56.8	56.0	61.2	58.5
FedProto	58.7	57.0	61.3	59.7
FedOGM	58.6	57.4	60.1	58.5
FedPMR	56.4	55.5	61.9	60.3
DivFL	57.4	55.6	61.1	58.6
FedMSplit	58.9	56.9	61.7	60.0
BMSFed	**62.4**	**59.8**	**63.5**	**60.9**

scales.

The relationship between modality selection and imbalance degree. In Eqs. 5.18 and 5.19, we use a conflict resolution strategy based on local imbalance ratio to realize balanced modality selection. We visualize the proportions of audio and visual modalities selected in each round and the global imbalance ratio. It is clear from Figure 5.4 that audio is the dominant modality (imbalance ratio is always greater than 1) and modality imbalance is gradually alleviated as training progresses (global imbalance ratio shows a downward trend). In addition, the proportion of selected visual modality follows the same trend (larger in the early stage of training and becomes smaller later), implying the rationality of our selection strategy based on imbalance ratio and its effectiveness on mitigating bias.

Effectiveness on modality incongruity scenario. All above experiments assume that each client initially has complete modal data. Here, we build the stimulation of modality incongruity scenario, in which half of the clients have data with two modalities, and the other half only have data in one modality: random audio or visual. The results are shown in Table 5.5 (The results of pow-d is not available because it is not applicable to this scenario). Our BMSFed still makes impressive improvement compared with all other baselines (by up to 3.5% on IID CREMA-D), illustrating the good generalization ability of our method in different scenarios. It is worth mentioning that FedMSplit performs better than before because it is specifically designed for modality incongruity.

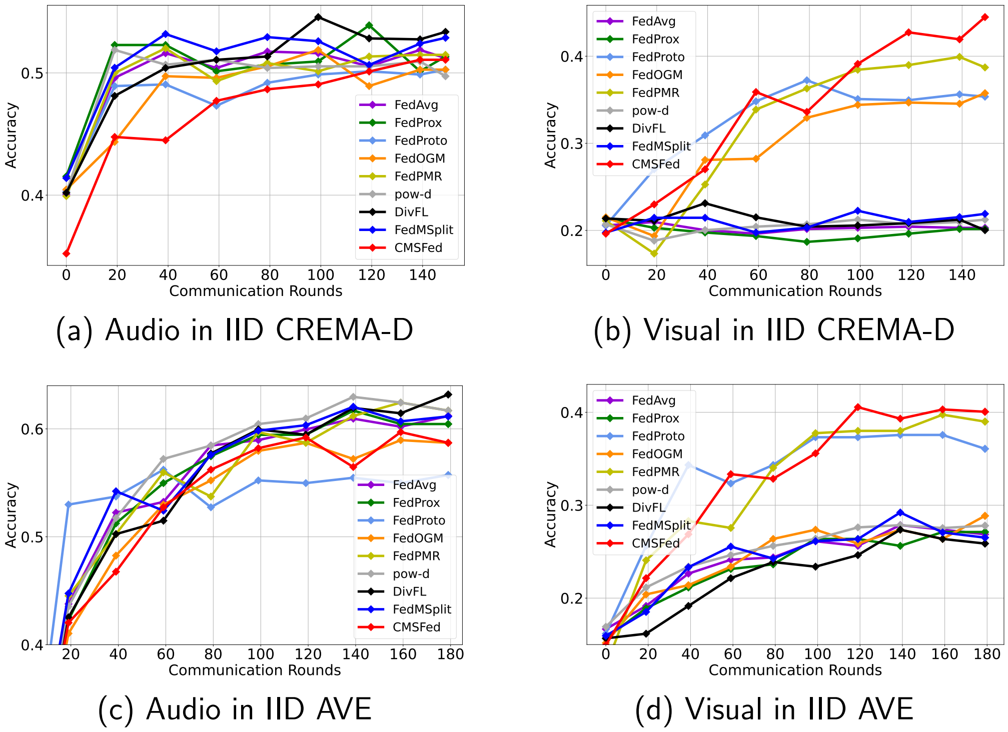
5.4.5 More Results

The unimodal performance. The accuracy curves of each modality of all baselines and our method on CREMA-D and AVE are shown Figs. 5.5 and 5.6.

The uni-modal results under non-IID settings are consistent with the observations under IID settings.

Comparison with more methods for modality imbalance. Here are more results comparing our BMSFed with more baselines for modality imbalance. AGM [64], G-blending [106] and Greedy [112] are all designed for modality imbalance problem in centralized scenario and we extend them to multi-modal FL settings. BMSFed still achieves the best performance on CREMA-D and AVE.

Comparison on Image-Text dataset with two sota FL methods for statistical heterogeneity. We evaluate our method on the image-text dataset CrisisMMD [1] to show its effectiveness on text modality. Moreover, we choose two more SOTA



(a) Audio in IID CREMA-D

(b) Visual in IID CREMA-D

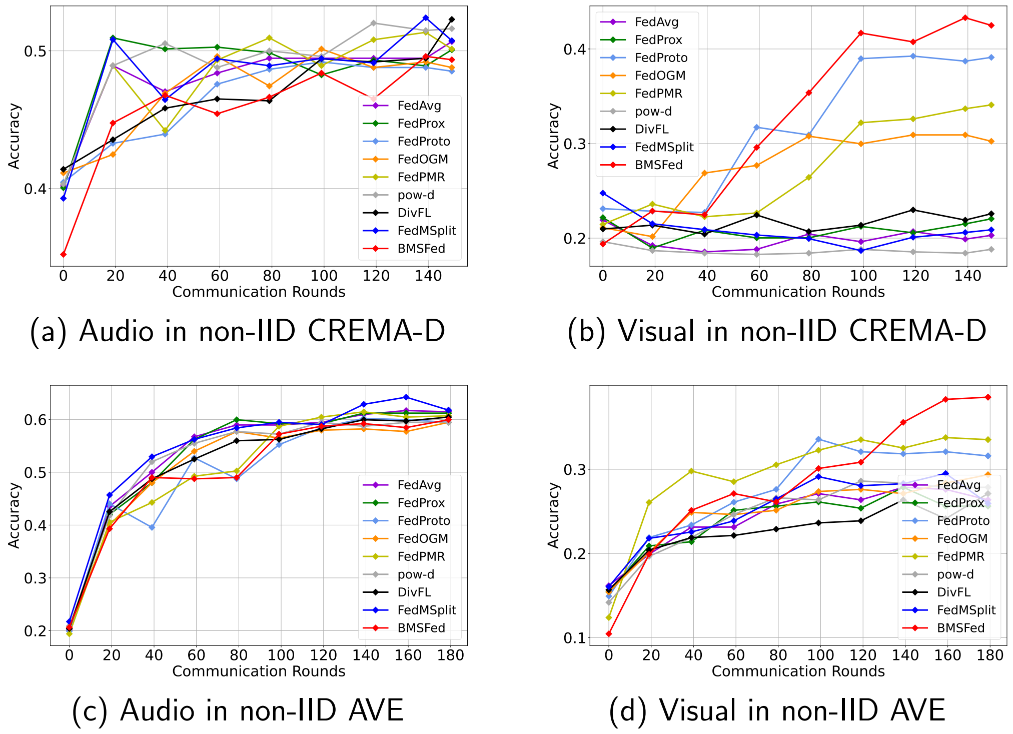
(c) Audio in IID AVE

(d) Visual in IID AVE

Figure 5.5: The performance of each modality compared with other baselines on CREMA-D and AVE under IID settings.

FL methods FedNH [20] and FedPAC [117] for statistical heterogeneity FedNH and FedPAC for comparison. The results are shown in table 5.7, our method still achieves the best.

More studies about claim and more settings. We fix the selection scheme for each client but differs among clients. As shown in table 5.8, this scheme is still better than FedAvg, proving “uni-modal local training may contribute more to global model”. But they are worse than randomly modality selection, because some data is never selected. The results of CREMA-D with more clients with different α are in table 5.9.



(a) Audio in non-IID CREMA-D

(b) Visual in non-IID CREMA-D

(c) Audio in non-IID AVE

(d) Visual in non-IID AVE

Figure 5.6: The performance of each modality compared with other baselines on CREMA-D and AVE under non-IID settings.

Table 5.6: Comparison with more modality imbalance methods. 'C' and 'A' denote CREMA-D and AVE respectively.

	C_{IID}	$C_{non-IID}$	A_{IID}	$A_{non-IID}$
AGM [64]	55.1	52.1	63.4	60.1
G-blending [106]	54.4	52.8	62.2	61.3
Greedy [112]	54.0	53.9	62.9	60.7
BMSFed	**64.5**	**61.6**	**64.7**	**62.1**

Table 5.7: Results compared with two sota FL methods on CREMA-D and an image-text dataset CrisisMMD.

	CREMA-D		CrisisMMD [1]	
	IID	non-IID	IID	non-IID
FedAvg	50.7	49.8	85.4	82.1
FedNH	58.6	56.3	87.3	85.8
FedPAC	59.7	56.4	87.1	86.5
FedPMR	55.5	55.1	86.6	86.0
DivFL	51.7	50.8	85.8	83.5
FedMSplit	52.4	51.6	85.9	84.7
BMSFed	**64.5**	**61.6**	**88.7**	**87.4**

Table 5.8: More results proving "uni-modal local training may contribute more to global model".

	FedAvg	-0.5	-0.8	-0.5-fix	-0.8-fix
A	51.2	50.5	48.1	49.6	45.2
V	20.6	34.6	50.9	34.1	44.8
A-V	50.7	55.7	61.2	55.0	57.3

Table 5.9: CREMA-D with more clients with different α.

	N=20,α=1	N=20,α=0.5	N=40,α=1	N=40,α=0.5
FedAvg	47.5	46.3	47.1	42.9
BMSFed	57.8	55.0	56.4	53.3

5.5 Remarks

In this thesis, we analyze traditional client selections and find their ineffectiveness in MFL. We further reveal that there exists strong modality-level bias due to the modality imbalance during the training iterations and uni-modal training on some clients may contribute more to the global model than multi-modal training. To address this issue, we propose the balanced modality selection scheme for MFL (BMSFed) with modality-level gradient decoupling to release the potential of all modalities and maximize the gradient diversities to improve global aggregation. We also introduce a modal enhancement loss to optimize the local update process. Our method does not introduce additional local training costs and communication overheads compared with previous methods. Extensive experiments on four datasets demonstrate the superiority of our method in performance and applicability under different modal combinations, data distributions and modality incongruity scenarios.

Chapter 6

Conclusions and Suggestions for Future Research

6.1 Work Summary

This thesis explores the challenges and solutions in maximizing the extraction and utilization of multimodal knowledge in various complex scenarios, addressing critical challenges in multimodal learning (MML), multimodal domain generalization (MMDG), and multimodal federated learning (MFL). Below is a summary of the key contributions and findings.

- A detached and interactive multimodal learning framework (DI-MML) is proposed to eliminate modality competition by independently training each modality with isolated objectives. A shared classifier and a dimension-decoupled unidirectional contrastive (DUC) loss enable cross-modal interaction, while an instance-level logit weighting strategy ensures balanced inference. Experiments demonstrate significant improvements in both unimodal and multimodal performance.

- To enhance multimodal domain generalization, a cross-modal representation flattening method (CMRF) is proposed, optimizing in representation space rather than parameter space to align flatness across modalities. By interpolating cross-modal representations and leveraging knowledge distillation, CMRF flattens high-loss regions and enhances weaker modalities. Adaptive weighting and supervised contrastive loss further improve generalization. Experiments on EPIC-Kitchens and HAC datasets validate its effectiveness in mitigating modality competition and improving generalization.

- For multimodal federated learning, a balanced modality selection framework (BMSFed) is introduced to address modality imbalance and global modal bias. It selects modalities rather than clients for training, using a modal enhancement loss with global prototypes and submodular optimization to ensure balanced learning. BMSFed outperforms baselines on multiple datasets, including CREMA-D, AVE, CG-MNIST, and ModelNet40, under both IID and non-IID settings, proving its robustness and scalability.

6.2 Future Work

The development of multimodal learning has opened the ages to solving tasks that previously could not be completely solved unimodally with multimodal data. Besides, with the development of large models in recent years, vast strong large language models (LLMs) as well as multimodal large language models (MLLMs) have emerged. They have shown much better performance than traditional small models on many downstream tasks. Therefore, how to use multimodal knowledge to solve more difficult tasks that are difficult to solve unimodally, and how to effectively use the knowledge of multimodal large models are extremely challenging. In the future, I plan to perform research on the following two directions:

First, although large models show superior performance, it is difficult to deploy them on local devices with limited computing resources. Therefore, extracting the specific knowledge from large models to small models allows the strengths of MLLM to be exploited at the edge. For example, when the local modal data is limited in quantity or poor in quality, knowledge can be extracted from large models of other modalities to promote local models.

Second, multimodal Retrieval-Augmented Generation (RAG) provides a promising direction for solving complex tasks by leveraging both retrieval capabilities and generative abilities across multiple modalities. We plan to explore how multimodal RAG can effectively integrate and utilize external multimodal knowledge bases to enhance the reasoning and generation capabilities of MLLMs. Specifically, I aim to investigate how to design efficient retrieval mechanisms that can handle diverse and heterogeneous data sources (e.g., text, images, videos) and align them with the generative model's input requirements.

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689, 2020.

[2] Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765, 2024.

[3] Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20208–20217, 2022.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022.

[6] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2123–2132, 2021.

[7] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[8] Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, and Jeff Bilmes. Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*, 2022.

[9] 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 transactions on affective computing*, 5(4):377–390, 2014.

[10] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[11] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[12] Chen Chen, Yuchen Hu, Qiang Zhang, Heqing Zou, Beier Zhu, and Eng Siong Chng. Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12607–12615, 2023.

[13] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[14] Jiayi Chen and Aidong Zhang. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 87–96, 2022.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

[17] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021.

[18] MMAction Contributors. Openmmlab's next generation video understanding toolbox and benchmark. 2020.

[19] Gerard Cornuejols, Marshall Fisher, and George L Nemhauser. On the uncapacitated location problem. In *Annals of Discrete Mathematics*, volume 1, pages 163–177. Elsevier, 1977.

[20] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7314–7322, 2023.

[21] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[22] Yongheng Deng, Feng Lyu, Ju Ren, Huaqing Wu, Yuezhi Zhou, Yaoxue Zhang, and Xuemin Shen. Auction: Automated and quality-aware client selection framework for efficient federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(8):1996–2009, 2021.

[23] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021.

[25] 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*, pages 8632–8656. PMLR, 2023.

[26] Jian-hui Duan, Wenzhong Li, Derun Zou, Ruichen Li, and Sanglu Lu. Federated learning with data-agnostic distribution fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8074–8083, 2023.

[27] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[28] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. Overcome modal bias in multi-modal federated learning via balanced modality selection.

[29] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. *arXiv preprint arXiv:2407.19514*, 2024.

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

[31] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Jiaqi Zhu, and Song Guo. Balanced multi-modal federated learning via cross-modal infiltration. *arXiv preprint arXiv:2401.00894*, 2023.

[32] Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. Stemm: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, 2022.

[33] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[34] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

[35] Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-specific learning rate control for multimodal classification. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pages 412–422. Springer, 2020.

[36] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

[37] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[38] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

[39] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

[40] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[41] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.

[42] Dong-Kyun Han and Ji-Hoon Jeong. Domain generalization for session-independent brain-computer interface. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–5. IEEE, 2021.

[43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[46] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[47] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

[48] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

[49] 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.

[50] Ahmed Imteaj, Urmish Thakker, Shiqiang Wang, Jian Li, and M Hadi Amini. A survey on federated learning for resource-constrained iot devices. *IEEE Internet of Things Journal*, 9(1):1–24, 2021.

[51] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

[52] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[53] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

[54] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[55] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[56] Ramandeep Kaur and Sandeep Kautish. Multimodal sentiment analysis: A survey and comparison. *Research anthology on implementing sentiment analysis across multiple disciplines*, pages 1846–1870, 2022.

[57] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[58] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[59] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[60] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[63] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

[64] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22214–22224, 2023.

[65] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[66] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[67] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[68] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[69] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2019.

[70] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[71] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023.

[72] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8635–8643, 2021.

[73] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021.

[74] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2218–2227, 2020.

[75] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[76] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[77] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020.

[78] Serkan Musellim, Dong-Kyun Han, Ji-Hoon Jeong, and Seong-Whan Lee. Prototype-based domain generalization framework for subject-independent

brain-computer interfaces. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 711–714. IEEE, 2022.

[79] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.

[80] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[81] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. Meta distribution alignment for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2022.

[82] Solmaz Niknam, Harpreet S Dhillon, and Jeffrey H Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine*, 58(6):46–51, 2020.

[83] Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[84] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[85] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022.

[86] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1807–1818, 2022.

[87] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[88] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[89] Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity. *IEEE Journal on Selected Areas in Information Theory*, 3(2):197–205, 2022.

[90] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[91] Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Domain generalization of 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18077–18087, 2023.

[92] Shashi Kant Shankar, Luis P Prieto, María Jesús Rodríguez-Triana, and Adolfo Ruiz-Calleja. A review of multimodal learning analytics architectures. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)*, pages 212–214. IEEE, 2018.

[93] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[94] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[95] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[96] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

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

[98] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8432–8440, 2022.

[99] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[100] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[101] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better

representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.

[102] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023.

[103] Haozhao Wang, Zhihao Qu, Song Guo, Ningqi Wang, Ruixuan Li, and Weihua Zhuang. Losp: Overlap synchronization parallel with local compensation for fast distributed training. *IEEE Journal on Selected Areas in Communications*, 39(8):2541–2557, 2021.

[104] Haozhao Wang, Wenchao Xu, Yunfeng Fan, Ruixuan Li, and Pan Zhou. Aocc-fl: Federated learning with aligned overlapping via calibrated compensation. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.

[105] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3769–3778, 2023.

[106] 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.

[107] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023.

[108] Yunqi Wang, Furui Liu, Zhitang Chen, Yik-Chung Wu, Jianye Hao, Guangyong Chen, and Pheng-Ann Heng. Contrastive-ace: Domain generalization through alignment of causal mechanisms. *IEEE Transactions on Image Processing*, 32:235–250, 2022.

[109] Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, and Changsheng Xu. Active exploration of multimodal complementarity for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6492–6502, 2023.

[110] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27338–27347, 2024.

[111] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210*, 2020.

[112] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

[113] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[114] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[115] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):537–547, 2020.

[116] Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022.

[117] Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023.

[118] Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2):1188–1200, 2020.

[119] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[120] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

[121] Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*, 2021.

[122] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022.

[123] Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. Few-shot classification with contrastive learning. In *European Conference on Computer Vision*, pages 293–309. Springer, 2022.

[124] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[125] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7097–7107, 2022.

[126] Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834, 2022.

[127] Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, and Peng Cui. Rethinking the evaluation protocol of domain generalization. *arXiv preprint arXiv:2305.15253*, 2023.

[128] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023.

[129] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[130] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[131] Qiulin Zhang, Zhuqing Jiang, Qishuo Lu, Zhengxin Zeng, Shang-Hua Gao, and Aidong Men. Split to be slim: an overlooked redundancy in vanilla convolution. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3195–3201, 2021.

[132] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5189–5202, 2023.

[133] Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, and Peng Cui. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20247–20257, 2023.

[134] Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on iot data. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 43–54. IEEE, 2022.

[135] Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *International Journal of Computer Vision*, 131(9):2377–2387, 2023.

[136] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020.

[137] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. Intra- and inter-modal curriculum for multimodal learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3724–3735, 2023.

[138] Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, and Ruixuan Li. Flatten long-range loss landscapes for cross-domain few-shot learning. *arXiv preprint arXiv:2403.00567*, 2024.