徽标, 公司名称

描述已自动生成

**2025年（第18届）**

**中国大学生计算机设计大赛**

人工智能实践赛作品报告

作品编号：　　　　　　2025039534

作品名称：　反诈专家——基于大模型的多模态风险内

　容鉴别器

填写日期：　　　　　　2025.04.16

**目 录**

[第1章 作品概述 1](#_Toc195729369)

[1.1 研究背景 1](#_Toc195729370)

[1.2 主要功能 1](#_Toc195729371)

[1.2.1 虚假文本检测 1](#_Toc195729372)

[1.2.2 虚假图文检测 2](#_Toc195729373)

[1.2.3 虚假视频检测 2](#_Toc195729374)

[1.3 创新点 2](#_Toc195729375)

[1.3.1 SDSA 数据去冗余与语义对齐机制 2](#_Toc195729376)

[1.3.2 MIB多模态平衡融合机制 2](#_Toc195729377)

[1.3.3 基于个性化调制器的LoRA大模型微调方法 2](#_Toc195729378)

[1.4 应用价值与推广前景 3](#_Toc195729379)

[第2章 问题分析 3](#_Toc195729380)

[2.1 问题来源 3](#_Toc195729381)

[2.2 现有解决方案 4](#_Toc195729382)

[2.2.1 数据去冗余与语义对齐 4](#_Toc195729383)

[2.2.2 多模态平衡融合 4](#_Toc195729384)

[2.2.3 联邦式多模态协同建模 4](#_Toc195729385)

[2.3 本作品要解决的痛点问题 5](#_Toc195729386)

[2.3.1 多模态数据冗余与语义不一致 5](#_Toc195729387)

[2.3.2 融合过程模态失衡 5](#_Toc195729388)

[2.3.3 分布式数据难以统一建模 5](#_Toc195729389)

[2.4 解决问题的思路 6](#_Toc195729390)

[第3章 技术方案 7](#_Toc195729391)

[3.1 系统架构 7](#_Toc195729392)

[3.2 基于语义蒸馏和结构相似性的三重对齐网络 7](#_Toc195729393)

[3.2.1 总体介绍 7](#_Toc195729394)

[3.2.2 技术方案：系统组件 8](#_Toc195729395)

[3.3 基于动态权重分配的多模态信息平衡技术 8](#_Toc195729396)

[3.4 基于个性化调制器的LoRA大模型微调方法 11](#_Toc195729397)

[第4章 系统实现 11](#_Toc195729398)

[4.1 数据来源与预处理流程 11](#_Toc195729399)

[4.2 多模态信息平衡融合机制的实现 12](#_Toc195729400)

[第5章 测试分析 14](#_Toc195729401)

[第6章 作品总结 16](#_Toc195729402)

[6.1 作品特色与创新点 16](#_Toc195729403)

[6.2 应用推广 16](#_Toc195729404)

[6.3 作品展望 17](#_Toc195729405)

[参考文献 18](#_Toc195729406)

# 作品概述

## 研究背景

随着通信技术与数字技术的迅猛发展与广泛应用，风险内容在网络空间中呈现出快速增长的趋势，成为当前社会管理与网络空间治理面临的重要挑战。多模态风险内容频繁出现于诈骗短信与电话、虚假新闻传播、伪造人脸图像与音视频等形式中。这些风险内容通常来源于多种模态媒介，包括文本、图像、音频、视频及流数据等，不仅形式多样、体量庞大，而且真假难辨，给识别与治理带来了显著难度。

在实际应用中，风险内容往往以混合模态出现，增加了传统识别技术的复杂性。因此，如何进行数据清洗与语义对齐、实现各模态的平衡融合、同时提取各模态数据中的关键特征并保留其语义信息，成为该领域亟待攻克的核心难题。为应对这一挑战，有必要围绕诈骗识别、伪造视频检测、虚假新闻识别等典型多模态风险识别任务，开展系统性的研究与技术应用，为网络空间治理提供有力支撑。

## 主要功能

本项目围绕多模态风险内容识别任务，充分融合了多项关键技术创新，旨在提升虚假内容检测的准确性、鲁棒性与实用性，设计并实现了以下三个核心功能：

### 虚假文本检测

系统通过多模态语义理解模型对文本中的潜在虚假风险信息进行识别，并输出带有风险点的摘要。用户可直观看到哪些句子或表达具有误导性或潜在风险，从而更好地理解文本中的虚假成分和风险来源。

### 虚假图文检测

系统利用图文对齐与特征异构融合机制，精准定位图像中可能存在篡改、合成、伪造的区域，并通过可视化方式以红色边框框选显示。用户在查看图像时即可直观地看到系统标记的“风险区域”，提升对图像真实性的判断能力。

### 虚假视频检测

系统对视频序列进行分析，识别深度伪造（如换脸）、剪辑合成等异常行为，并同样采用“伪造区域框选”方式呈现。系统会在视频内容中以动态图形式标注出异常区域，辅助用户快速定位和理解视频中的虚假内容。

## 创新点

### SDSA 数据去冗余与语义对齐机制

针对当前多模态数据在采集过程中普遍存在数据冗余严重、噪声比例高、语义不一致等问题，项目提出SDSA（Selective De-redundancy and Semantic Alignment）数据去冗余与语义对齐机制，该机制为虚假内容识别提供了更高质量、更一致性的输入数据，有效提高了模型训练效率与识别准确率。

### MIB多模态平衡融合机制

针对多模态信息融合过程中“强模态主导、弱模态退化”等常见问题，项目设计并实现了BMLHF（Balanced Multi-modal Learning with Hybrid Fusion）多模态平衡融合机制，该机制显著提升了模型对多模态输入的综合理解与决策能力，在应对复杂场景下的虚假图文与视频识别任务中展现出卓越的适配性。

### 基于个性化调制器的LoRA大模型微调方法

为提升模型在数据分布广泛、隐私保护要求高的实际环境下的可部署性与泛化能力，项目提出FEDPM（Federated Learning based Personalized Modulators​ for Multi-LLMs）基于个性化调制器的LoRA大模型微调方法，该方法突破了传统多模态建模对集中数据依赖的限制，在保障隐私的同时，显著提升了模型的跨域泛化能力与工程可落地性。

## 应用价值与推广前景

本项目聚焦虚假文本、图文及视频内容识别，构建了多模态面向风险内容的鉴别体系，具有实际应用价值和推广前景。项目成果可全方面应用于网络安全治理、内容审核、媒体舆情监测等战略性领域，助力政府监管部门提高对虚假信息的识别效率与干预能力，强化社会治理数字化水平。同时，可为社交平台、新闻客户端、视频平台等内容发布方提供高精度内容合规审核工具，提升内容合规水平与用户信任度。

项目采用模块化架构，支持多系统集成，具备良好的可部署性与扩展性；联邦学习式架构保障数据安全合规，适用于多组织、跨行业的协同应用场景。随着AIGC等生成内容的快速发展和虚假信息传播手段不断演化，社会各界对多模态识别技术的需求日益增长，项目具备良好的技术引领性和商业转化潜力，推广前景广阔。

# 问题分析

## 问题来源

随着大模型技术的快速发展和网络数据的爆炸式增长，文本、图像、视频等多模态信息在互联网中高速传播。与此同时，虚假文本、伪造图像与深度伪造视频等风险内容不断涌现，形成更具欺骗性和传播力的多模态虚假信息，严重威胁网络安全与社会信任。

当前主流的风险识别技术仍以单模态或浅层融合为主，难以处理多模态数据间的语义对齐、模态失衡和跨源异构等复杂问题，且普遍缺乏在高质量数据清洗与隐私合规方面的系统支持，识别效果受限、泛化能力不足。

因此，亟需构建面向多模态风险内容的大模型识别技术体系，从数据预处理、特征融合到协同建模形成全流程能力，突破当前识别能力瓶颈，满足复杂风险场景下的智能识别需求。

## 现有解决方案

近年来，多模态风险内容检测逐渐成为人工智能与网络安全领域的研究热点，相关研究主要集中在数据清洗与去冗余、模态融合策略、多源建模机制等方向。然而，现有技术仍存在诸多局限，难以满足实际复杂场景中对高质量数据处理、模态平衡理解及隐私合规建模的综合需求。

### 数据去冗余与语义对齐

EANN[1]引入了一个事件鉴别器来减少事件差异。SpotFake[2]直接执行真/假新闻分类，无需重新依赖子任务。BDANN[3]采用带有域分类器的域自适应神经网络来减少事件差异。MVAE[4]引入了变分自动编码器这项工作得到了国家自然科学基金（No. 62076139）、南京邮电大学 1311 人才计划的支持。获取多模态特征。MEAN[5]在事件判别器和分类器中引入了交互机制。CAFE[6]提出了一种用于多模态假新闻检测的跨模态模糊性学习问题。MRML[7]通过度量学习来建立模态内的语义关系。另一组方法，如参考文献[8][9][10]所述，利用模态之间的语义相似性来帮助特征学习。例如，CAFE[7]确定文本和图像是否表现出相似性，将大多数不匹配的情况归类为假新闻。然而，这些假新闻检测方法无法有效去除冗余信息[11][12][13]并忽略新闻中的结构信息[14]。

### 多模态平衡融合

目前，解决模态不平衡问题的方法主要集中在音视频分类领域。其中，流行的做法包括单峰辅助[15]、梯度混合[16]和样本级模态评估[17]。例如，Du et al.[15]在训练有素的单模态模型的帮助下加强了多模态模型。Wang等[16]采用梯度混合来获得模态的最佳混合。Wei等[15]使用样本级模态评价指标增强了低贡献模态。然而，这些方法只关注解决音视频分类问题，并没有考虑 MFND 中的模态不平衡。新闻数据在特征差异、模态一致性和融合策略方面与音视频数据存在显著差异，使得这些不平衡的多模态学习方法不能直接用于 MFND。

### 联邦式多模态协同建模

当前在多语言大型语言模型（LLMs）领域的研究大多集中于构建统一的全局模型（如 Iacob 等，2025[18]；Weller 等，2022[19]；Ye 等，2024[20]），但这种方法往往受到所谓“多语言性诅咒”（multilinguality curse）和“负迁移”（negative transfer）问题的制约（Conneau 等，2020[21]；Wang 等，2020b[22]），即在不同语言之间共享参数时，由于语料分布的显著差异，反而会导致模型在特定语言上的性能下降。全局模型难以有效适应多语言之间的异质性，尤其在低资源语言场景中表现较差。

为缓解这一问题，一些研究引入了个性化联邦学习（personalized FL）策略，尝试在每个客户端上对全局模型进行微调（如 Jiang 等，2019[23]），但此类方法仍需手动设计微调模块的结构与超参数，如 LoRA 模块的插入位置、秩设定和学习率等，这不仅增加了部署成本，也易在低数据场景中出现过拟合。此外，如 FedL2P（Lee 等，2023[24]）等方法虽然通过联邦机制实现了个性化学习率或批归一化参数的调整，但其个性化能力较弱，难以在客户端差异较大的环境下实现有效泛化。

## 本作品要解决的痛点问题

### 多模态数据冗余与语义不对齐

在多模态虚假内容识别任务中，数据往往来源复杂、模态多样，尤其在网络环境下采集的图像、文本、音频、视频等信息中，存在大量冗余片段、模糊标签和语义不对齐的问题。这不仅降低了训练数据的有效性，还可能引入误导性特征，影响模型的学习效率和识别准确率。现有方法多依赖单模态静态清洗策略，难以对多模态语义关系进行统一建模与优化。项目提出 SDSA 数据去冗余与语义对齐机制，从多模态语义一致性的角度出发，对输入数据进行筛选、提纯与对齐，为模型提供更高质量的输入支撑，有效缓解该问题。

### 融合过程模态失衡

当前多模态融合方法普遍存在模态主导性强弱不均的现象。在图文、音视频等融合场景中，主导模态（如图像或视频）在训练过程中会压制弱模态（如文本或语音）的表达能力，导致弱模态特征被忽视，信息融合不充分，最终影响模型的鲁棒性和泛化能力。为应对这一问题，项目设计并实现了MIB多模态平衡融合机制，引入模态重要性动态调节策略和多路径融合结构，实现各模态之间的协同学习和语义互补，从而提升模型对复杂场景下多模态虚假内容的识别效果。

### 分布式数据难以统一建模

在多语言大模型的应用场景中，训练数据往往呈现出高度分布式和异质性特征，即不同语言、地区和用户群体的数据分布存在显著差异。这种差异导致传统集中式训练方法难以有效适应所有语言和客户端，统一建模效果显著受限。尤其在涉及低资源语言和隐私保护场景时，数据无法汇集至中心服务器进行统一训练，进一步加剧了模型性能退化和泛化能力不足的问题。因此，亟需一种能够在数据不出本地的前提下，有效实现模型个性化调优与结构适配的解决方案，以提升多语言模型在联邦环境下的可用性与公平性。

## 解决问题的思路

本项目旨在通过构建融合SDSA、BMLHF与FEDPM机制的多模态风险内容识别系统，提升模型在多源异构数据环境下的识别精度、处理效率及跨场景适应能力。系统面向**文本、图像与视频**共三种模态，实现虚假文本检测、图文一致性识别与伪造视频识别等核心功能。

在功能与性能层面，本系统在真实环境模拟测试中表现出良好的计算效率和识别能力。基于100次运行测试，模型平均响应时间为175ms，整体内存开销控制在边缘部署可接受范围内，适合联邦学习等分布式场景下的实际应用。在风险内容识别方面，系统在多个标准测试集上实现了96%的准确率，F1分数达到0.83，充分体现了其在精度与鲁棒性方面的优势。

在实验数据方面，项目选用两个常用的多模态虚假信息检测公开数据集：Weibo 和 Fakeddit，覆盖中英文社交平台、多领域事件，具有良好的代表性与挑战性。

表 2. 多模态虚假信息检测公开数据集信息

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **数据集名称** | **语言** | **样本规模** | **模态类型** | **数据特点** |
| **Weibo** | 中文 | 4700条真实新闻，4000条假新闻 | 文本+图像 | 来源于新浪微博，领域广泛，中文多模态代表语料 |
| **Fakeddit** | 英文 | 约60,000条标注样本 | 文本+图像 | Reddit平台采集，规模大、语义复杂 |

# 技术方案

## 系统架构

## 基于语义蒸馏和结构相似性的三重对齐网络

### 总体介绍

多模态信息冗余和结构信息丢失问题本质上是跨模态交互中因噪声干扰和特征对齐失准引发的有效信息筛选与语义结构完整性失衡。为解决这一问题，我们提出基于语义蒸馏和结构相似性的三重对齐网络（以下简称SDSA），通过分层语义蒸馏和跨模态相似性度量，动态对齐多模态表征，自适应过滤冗余信息以保留结构特征，系统分为两大部分：

1. **语义蒸馏**：通过蒸馏模态特定的特征，保留与假新闻检测任务相关的语义信息，同时消除冗余信息。
2. **结构对齐**：确保在模态内、模态间以及跨模态保留结构信息（邻域关系）。

整体架构的框架图如图3.1所示：

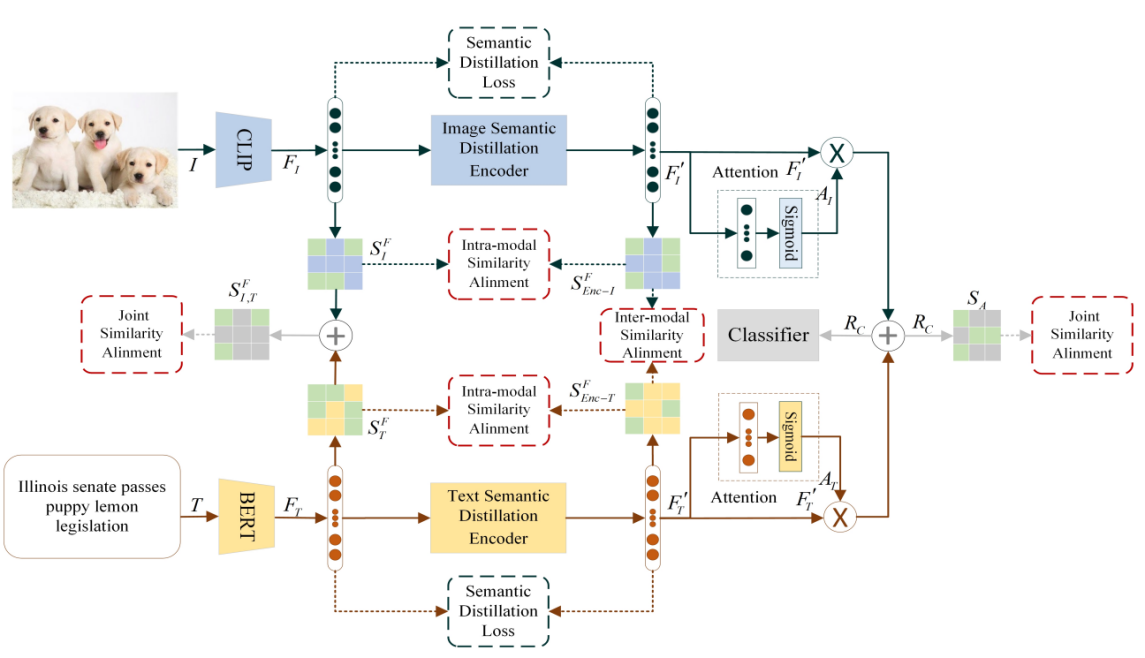


图3.1 SDSA方法总体架构图

### 技术方案：系统组件

3.2.2.1语义蒸馏

对于假新闻检测任务，语义信息由两部分组成：与任务相关的语义信息和与任务无关的语义信息（冗余信息）。对于，我们可以将其表示如下：

其中，代表与任务相关的信息，表示冗余信息。我们的目标是仅保留与任务相关的语义信息，并从中剔除冗余信息。这意味着是的充分表示。因此，我们的目的是最小化与之间的差异：

为便于计算，我们将上式等价转换为最小化条件熵和：

其中。如果和之间的 KL散度等于 0，那么就是的充分表示：

其中和表示预测分布，DKL表示 KL散度。因此，通过最小化和之间的 KL散度，我们确保只保留与任务相关的语义信息。

其中表示信息蒸馏编码器的参数。

3.2.2.2三重相似度对齐

为了保留原始数据的结构信息并减少跨模态差异，我们使用相似性矩阵来表示数据集的邻域结构，并提出了一个三重相似性对齐模块，以利用模态内和跨模态邻域结构中的相似性信息，该模块包括模态内相似性对齐、模态间相似性对齐和联合相似性对齐。

具体来说，给定一批输入实例，我们利用范数对进行归一化处理，得到，并为计算余弦相似性矩阵，该矩阵用于描述不同模态中的原始邻域结构。类似地，通过范数对编码特征进行归一化处理后，我们可以计算余弦相似性矩阵，它用于描述经过编码后的不同模态的邻域结构。

然后，这些相似性矩阵被用于优化。此外，跨模态融合特征通过范数进行归一化处理，得到，我们计算余弦相似性矩阵，以描述融合特征的邻域结构。

**模态内相似性对齐：**为了保留每个模态内原始特征的邻域结构，我们将与进行对齐，并计算和之间的均方误差损失。通过最小化这些模态内均方误差损失，我们可以尽可能多地保留模态内的邻域结构。我们将模态内相似性损失定义如下：

**模态间相似性对齐：**为了有效减少模态之间的差异，我们对齐和，计算与之间的均方误差损失。通过最小化这种模态间均方误差损失，我们旨在拉近跨模态的特征邻域结构，从而减少模态间的差异。模态间相似性损失函数如下：

**联合相似性对齐：**我们利用模态间邻域结构之间的互补关系来构建联合相似性矩阵 。我们将与进行对齐，以使融合特征能够保留来自各种模态的邻域结构。由于不同模态的语义丰富程度不同，我们引入一个超参数来调整联合相似性矩阵中不同模态结构信息的权重。这有助于更好地利用来自不同模态的邻域信息。构建方法定义如下：

因此，我们设计了一种三重相似性对齐方法，涵盖了模态内、模态间和联合相似性对齐。我们将整体相似性对齐损失定义如下：

## 基于动态权重分配的多模态信息平衡技术

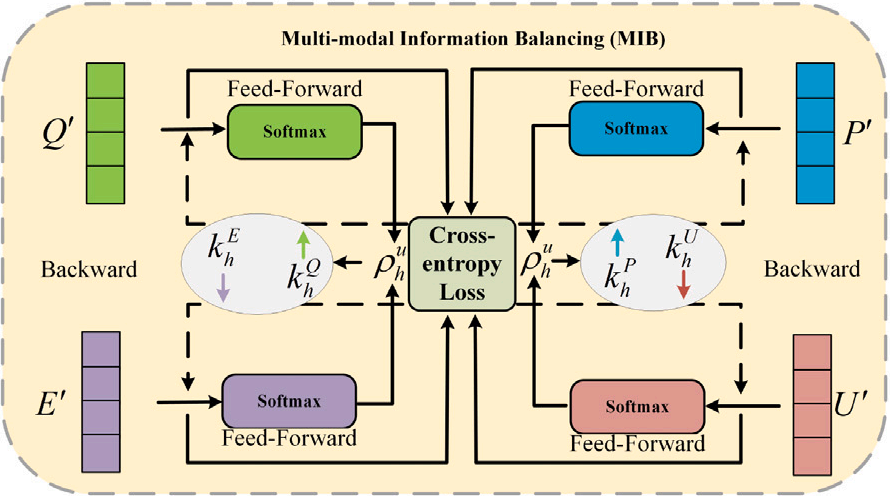


图3.1 多模态信息平衡模块架构图

模态不平衡问题是指在优化过程中，主导模态会抑制其他模态性能的现象。为解决这一问题，我们设计了多模态信息平衡（MIB）模块，在各模态的优化过程中动态地为它们分配权重，自适应地调节模型的优化过程，以实现模态平衡，如图3.1所示。

对于的每个通道，其参数为，梯度更新过程由下式表示：

其中是第次迭代更新后的参数。

我们通过更精细地监测信息，自适应地调整每个模态的梯度。这些信息可以表述为每个模态的特定信息，对数几率得分（logit score）能直接反映不同模态的激活程度。因此，在本文中，我们计算每个模态的模式视图和语义视图的对数几率得分，以探究每个模态的信息量。对数几率得分包含，其中和是的参数，表示中的第个特征，是第次迭代中的一个随机批次。我们定义来进一步量化不同模态视图的信息：

其中 ，是类别数。

我们设计差异率来衡量各模态对优化过程的影响：

越大，与其他模态视图相比，相应模态视图中的信息量就越大，从而导致模态不平衡。

信息量较大的模态在模型的优化过程中起着主导作用[7][8][9]，并抑制其他模态的优化过程。因此，我们设计平衡因子来平衡各模态的优化过程。

然后，我们将整合到公式 (1) 中。的更新过程如下：

通过使用，我们抑制了性能较好的模态视图的优化，而性能较差的模态视图则不受影响。

在获取多模态的融合特征后，我们将其输入到一个分类器中，该分类器包含一层多层感知器（MLP）和 激活函数，以获得预测标签：

其中是分类器，是分类器的参数。

为了增强模型的分类能力，我们最小化交叉熵损失：

为了增强特征的判别能力，我们采用跨模态监督对比损失。是中与属于同一类别的索引集，是中与类别不同的索引集。

其中，是的基数，是一个标量参数，根据经验设置为。因此，总损失可以表示为：

其中，是一个平衡因子。

我们使用 Adam 优化器在总损失下寻找最优的网络参数。对于测试样本集 ，我们使用最优参数、、、、、来获取相应的特征，然后将这些特征输入到分类器中以获得标签。通过我们的方法，可以调节每个模态的优化过程，有效缓解模态不平衡问题。

## 基于个性化调制器的LoRA大模型微调方法

# 系统实现

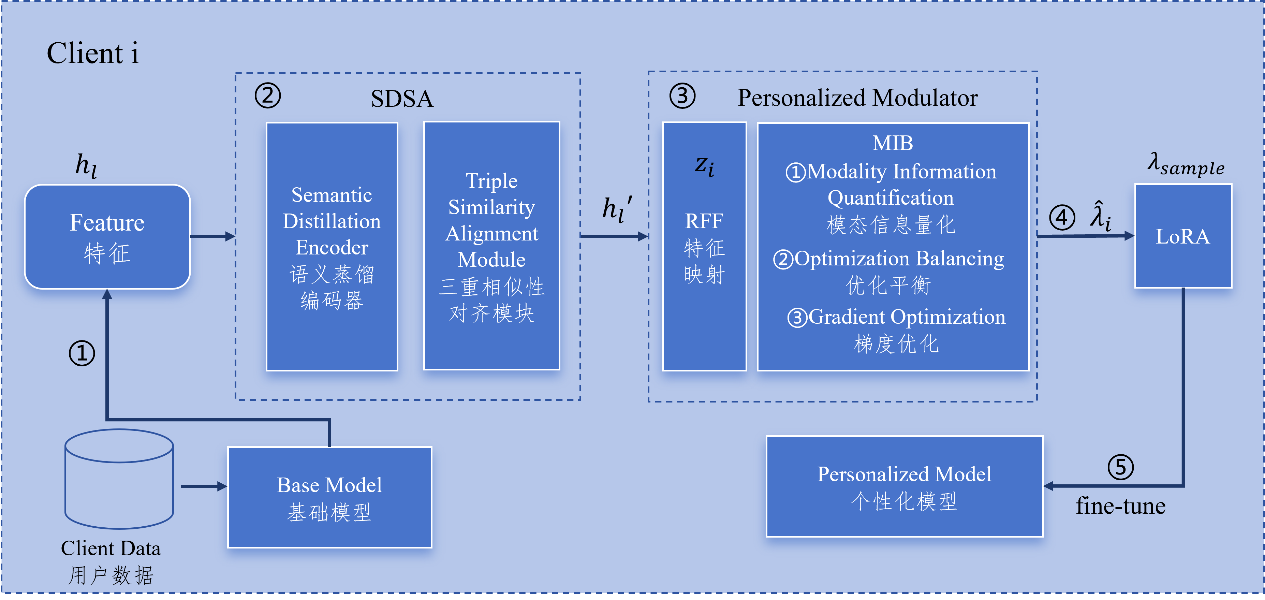


图 4.1 系统框架图

本项目围绕多模态风险内容识别的关键问题，依托SDSA去冗余与语义对齐机制、MIB多模态平衡融合机制与FEDPM联邦双通道学习框架三大核心技术，在多模态数据处理、模型构建、系统部署与优化等方面展开了系统性的工程实现与迭代优化，构建了具有实用价值与工程可落地性的风险内容智能识别系统。

## 数据来源与预处理流程

为验证SDSA去冗余与语义对齐机制，本系统选取两个典型多模态虚假内容数据集作为基础训练与测试资源，分别为中文社交平台微博上的**Weibo 数据集**和英文社交平台Reddit构建的 **Fakeddit 数据集**。在微博数据集中，训练集包含7532条新闻，其中3749条为假新闻，3783条为非假新闻，测试集包含1996篇新闻文章。在Fakeddit数据集中，我们从Fakeddit训练集中随机选择30,000对图像-文本对作为我们的训练集，从测试集中随机选择10,000对图像-文本对作为我们的测试集。实验部署在带有PyTorch的NVIDIAGeFor-ce1080TiGPU上。

我们的基线模型包括模态特定特征提取器，具有三个完全连接层的编码器，注意网络和分类器。我们将此基线模型称为SDSA\*。SDSA\*+S表示在基线模型上增加三重相似性对齐模块。SDSA\*+I表示用语义蒸馏编码器替换基线模型的编码器。SDSA表示在基线模型中同时添加三重相似性模块和语义蒸馏编码器。我们在表5.1的底部部分报告了结果。

在数据处理方面，首先对原始新闻文本进行标准化预处理，包括去除多余符号及统一编码格式。随后采用与预训练模型（如 mBERT 或 XLM-R）配套的分词器进行编码。所有数据均按 8:1:1 的比例划分为训练集、验证集和测试集，并在个性化微调阶段仅使用本地数据进行训练。训练的具体流程如图所示：

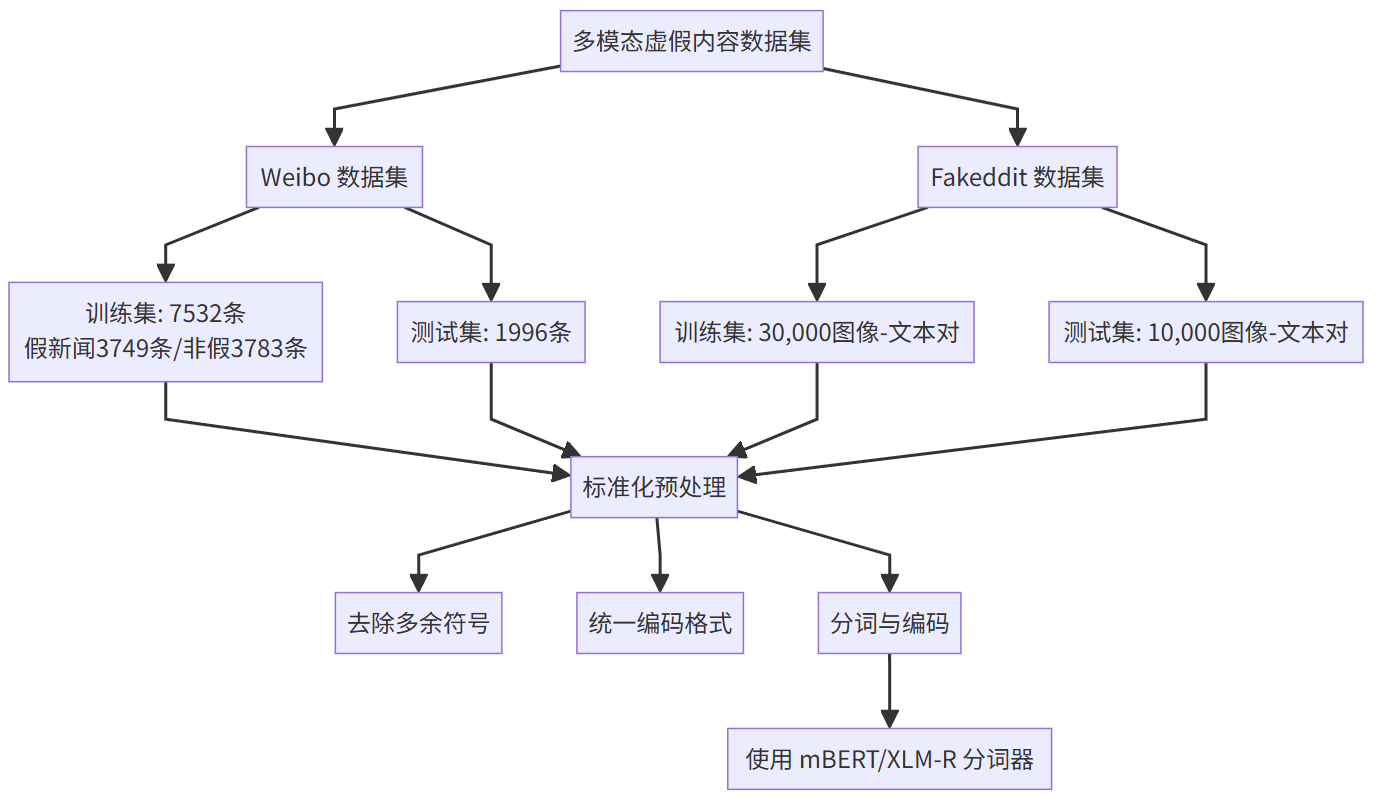


图 4. 多模态虚假内容数据集处理流程

## 数据有效性验证

为验证所提出 BMLHF 模型在多模态假新闻检测任务中的通用性与鲁棒性，本研究选取了两个具有代表性的数据集——Weibo（中文社交平台数据）与 Fakeddit（大规模英文多模态数据）进行综合评估。两者在语言类型、数据规模、内容结构及模态配比上具有显著差异，能够有效用于检验模型在多语言、多源异构场景下的适应能力。

在 Weibo 数据集上，BMLHF 模型在准确率和 F1 值方面均取得显著提升，完整模型在该数据集上达到 91.49% 的准确率和 91.14% 的 F1 值，相较于基础模型提升超 3 个百分点。这充分说明模型所提出的模态平衡机制和层次融合策略对中文社交文本与配图之间的语义耦合具有良好的建模能力，能够有效缓解因图像噪声或模态差异所引起的性能波动。

在 Fakeddit 数据集上，面对更大规模、更复杂的多模态数据输入（包含大量图文组合、跨主题新闻），BMLHF 依然保持稳定优势。实验表明，所提出模型在该数据集上同样优于所有对比基线方法，且在准确率、精确率和召回率等多项指标上均取得领先。这一结果验证了模型在大规模英文舆情数据中具有良好的可扩展性和泛化能力。

综上，Weibo 与 Fakeddit 两个数据集在语言、规模和结构上的互补性，使得实验具备广泛代表性。BMLHF 在两个数据集上的稳定表现进一步证明了其在多语言多模态假新闻检测任务中的有效性和通用性。

## 基于语义蒸馏和结构相似性的三重对齐网络的实现

本项目在系统实现过程中，首先构建了一个面向多模态虚假信息检测任务的三重对齐网络，结合语义蒸馏机制与结构相似性优化策略，以提升图文模态间的语义协同理解能力。具体而言，系统利用CLIP和BERT模型分别提取图像与文本的基础特征，并在此基础上引入语义蒸馏机制，通过轻量化编码器对特征进行语义精炼，增强其判别能力，并引入蒸馏损失以保障知识迁移效果。同时，为实现更精细的模态对齐，引入结构相似性三重损失策略，从模态内一致性、模态间对齐性以及融合空间结构保留三方面联合约束，优化特征表征质量。这一过程的训练伪代码如下：

| 算法1 基于语义蒸馏和结构相似性的三重对齐网络 |
| --- |
| 输入数据：图文对集合，图像模态和文本模态的特征矩阵和。  for each epoch：   1. 使用CLIP和BERT提取图像和文本基础特征、； 2. 对图像和文本基础特征、进行语义蒸馏编码，得到、； 3. 计算语义蒸馏损失； 4. 使用注意力子网融合、，得到融合特征； 5. 计算相似性对齐损失； 6. 使用分类器对融合特征进行分类，得到预测结果； 7. 反向传播更新参数；   输出：分类得到预测标签，计算预测准确率和F1分数等评估指标。 |

## 多模态信息平衡融合机制的实现

在多模态融合模块的实现过程中，项目引入了 MIB（Modal Information Bottleneck）平衡机制，该机制在每一轮训练中对不同模态的特征表达进行动态加权，有效解决了主导模态抑制其他模态表达的问题。通过差异率计算与平衡因子的引入，系统能够动态适配不同模态在融合过程中的权重比例，使得多模态信息得以充分、均衡地利用。其核心处理流程如下伪代码所示：

|  |
| --- |
| 算法2 基于动态权重分配的多模态信息平衡技术 |
| 输入数据：训练集 ，包含文本模态和图像模态以及标签矩阵 ；测试集 ，包含文本模态 和图像模态 。  for each epoch ：   1. 计算各模态特征 ； 2. 计算每个模态的对数几率得分 ； 3. 计算差异率 ； 4. 计算平衡因子 ； 5. 融合多模态特征 ； 6. 通过分类器 得到预测标签 ； 7. 反向传播更新参数；   optimizer.zero\_grad()  L\_total.backward()  更新模型参数；  输出：预测标签。 |

该方法的关键优势在于通过自适应的方式实现特征融合和模型参数更新，提升了模型的鲁棒性和对多模态信息的综合理解能力，为风险内容识别提供了更优的决策依据。

# 测试分析

表 5. 不同模型在Weibo和Fakeddit数据集上的比较结果[6]

| Method | Weibo | | | | | | | Fakeddit | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Acc | Rumor | | | Non-rumor | | | Acc | Rumor | | | Non-rumor | | |
| P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| EANN [KDD-2018] | 0.782 | 0.827 | 0.697 | 0.756 | 0.752 | 0.863 | 0.804 | 0.724 | 0.727 | 0.719 | 0.723 | 0.722 | 0.729 | 0.726 |
| SpotFake[BigMM-2019] | 0.892 | 0.902 | 0.964 | 0.932 | 0.847 | 0.656 | 0.739 | 0.819 | 0.801 | 0.848 | 0.824 | 0.839 | 0.790 | 0.813 |
| BDANN[IJCNN-2020] | 0.842 | 0.830 | 0.870 | 0.850 | 0.850 | 0.820 | 0.830 | 0.812 | 0.836 | 0.776 | 0.805 | 0.791 | 0.847 | 0.818 |
| HMCAN [SIGIR-2021] | 0.885 | 0.920 | 0.845 | 0.881 | 0.856 | 0.926 | 0.890 | 0.881 | 0.880 | 0.882 | 0.881 | 0.882 | 0.880 | 0.881 |
| MEAN [SPR-2022] | 0.894 | 0.900 | 0.870 | 0.890 | 0.890 | 0.910 | 0.90 | 0.910 | 0.930 | 0.890 | 0.910 | 0.890 | 0.930 | 0.910 |
| CAFE [WWW-2022] | 0.840 | 0.855 | 0.830 | 0.842 | 0.825 | 0.851 | 0.837 | 0.912 | 0.946 | 0.886 | 0.959 | 0.878 | 0.942 | 0.909 |
| MRML [ICASSP-2023] | 0.897 | 0.898 | 0.887 | 0.892 | 0.896 | 0.905 | 0.901 | 0.840 | 0.819 | 0.874 | 0.846 | 0.865 | 0.807 | 0.835 |
| SDSA\* | 0.898 | 0.916 | 0.888 | 0.902 | 0.880 | 0.909 | 0.894 | 0.941 | 0.946 | 0.936 | 0.941 | 0.936 | 0.947 | 0.942 |
| SDSA\*+S | 0.905 | 0.919 | 0.899 | 0.909 | 0.891 | 0.912 | 0.901 | 0.944 | 0.952 | 0.935 | 0.943 | 0.936 | 0.953 | 0.944 |
| SDSA\*+I | 0.906 | 0.924 | 0.895 | 0.909 | 0.887 | 0.919 | 0.903 | 0.950 | 0.946 | 0.954 | 0.950 | 0.954 | 0.946 | 0.950 |
| SDSA | 0.918 | 0.939 | 0.902 | 0.920 | 0.896 | 0.935 | 0.915 | 0.953 | 0.965 | 0.939 | 0.952 | 0.940 | 0.966 | 0.953 |

从表5.1可以看出，与SDSA\*相比，SDSA\*+S在微博数据集和Fakeddit数据集上分别提⾼了0.8%和0.3%。SDSA\*+I在两个数据集上也分别提⾼了0.9%和0.9%。SDSA\*在两个数据集上分别提⾼了2.2%和1.3%。这些实验结果表明了我们的语义蒸馏编码器和三重相似性对齐模块的有效性。

为验证FEDPM联邦双通道学习框架的研究围绕“已见客户端（seen clients）”情境下的模型个性化能力展开评估，并重点对比了 LoRA、FedL2P 等代表性参数高效微调方法的性能表现。从表5.2中的对比结果可以看出，LoRA 方法在多数语言上取得了较为稳定的准确率表现，如在英语（eng）、皮钦语（pcm）等高资源语言中准确率均在 90% 以上，但在如阿姆哈拉语（amh）等低资源语言上准确率仅为 45.74%。相比之下，FedL2P 方法在整体准确率上略有提升，尤其在 run（81.99%）、som（61.00%）等中等资源语言上相较于 LoRA 有明显进步。

然而，从“wins”列统计的最佳表现次数来看，LoRA 与 FedL2P 在 16 种语言中均未取得明显优势，显示出手工配置的个性化策略仍存在适应性不足的问题，尤其在多语言迁移和极低资源客户端中，个性化方法的差异化效果并不显著。这为后续自动策略生成机制的引入提供了实践动因。

表 5. MasakhaNEWS 设置中各语言在已见客户端上的准确率（Mean ± SD）[10]

| r | eng | som | run | fra | lin | ibo | amh | hau | pcm | swa | Wins |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LoRA | 90.44±0.10 | 60.09±0.32 | 81.37±0.51 | 88.63±0.00 | 83.53±0.54 | 79.83±0.24 | 45.74±0.00 | 75.79±0.00 | 96.05±0.00 | 78.99±0.00 | 0 |
| FedL2P | 90.72±0.59 | 61.00±1.16 | 81.99±0.88 | 89.10±0.67 | 83.91±0.00 | 79.66±0.24 | 45.74±0.00 | 76.73±1.12 | 96.05±0.00 | 79.69±0.40 | 0 |

综上，MasakhaNEWS 数据集的多语种分布与数据划分方式充分模拟了真实世界中的非独立、多样性数据环境；而从实验结果来看，现有的参数高效个性化方法虽具一定效果，但在低资源与语言迁移能力上仍存在优化空间。这也进一步说明自动化、稀疏化策略学习机制（如 FedP2EFT）在提升联邦个性化微调能力方面具有现实意义和研究价值。

表 5.3 BMLHF各模块配置在微博和Fakeddit数据集上的性能对比[11]

| **Model** | **Twitter Acc (%)** | **Twitter F1 (%)** | **Weibo Acc (%)** | **Weibo F1 (%)** |
| --- | --- | --- | --- | --- |
| **Base** | 90.12 | 89.95 | 88.29 | 88.15 |
| **Base + HF** | 93.02 | 92.81 | 89.91 | 89.67 |
| **Base + MIB** | 92.37 | 92.20 | 90.58 | 90.35 |
| **BMLHF (Full)** | 94.32 | 94.17 | 91.49 | 91.14 |

从表中结果可以看出，基础模型在 Twitter 和 Weibo 数据集上的准确率分别为 90.12% 与 88.29%，F1 分数分别为 89.95% 与 88.15%，为本研究中最低配置的性能基准。进一步在此基础上加入 HF 模块后，模型性能显著提升，Twitter 上准确率上升至 93.02%，Weibo 上提升至 89.91%，表明层次融合机制在强化模态间协同方面具有良好效果。

而引入 MIB 模块的配置（Base + MIB）同样在两个数据集上表现出稳定增益，准确率分别达到 92.37%（Twitter）与 90.58%（Weibo），说明模态信息平衡机制能够有效缓解“强模态主导”问题，促进弱模态信息的有效利用。

最终，完整模型 BMLHF 在两个数据集上均取得最高准确率和 F1 值，分别为 94.32% / 94.17%（Twitter）和 91.49% / 91.14%（Weibo），显著优于所有消融模型，验证了所提多模态协同机制与模态平衡机制的互补性与协同增益效应。

# 作品总结

## 作品特色与创新点

本项目聚焦于多模态风险内容识别的关键技术难题，结合大模型的深度语义理解能力与跨模态建模优势，系统提出三项核心创新机制，构建了具备强泛化性、高效率、低成本的多模态风险内容识别方案，具有鲜明的技术特色和理论创新。

**1.SDSA数据去冗余与语义对齐机制**：针对多模态数据中普遍存在的冗余信息、噪声干扰与语义不一致问题，提出SDSA机制，从特征层对数据进行选择性清洗与语义对齐，确保输入数据的纯净性与模态协同性，为后续建模提供更具代表性的训练样本，提升识别准确率和模型鲁棒性。

**2.MIB多模态平衡融合机制**：在多模态融合任务中常见的主模态主导、弱模态退化问题基础上，项目设计并实现BMLHF机制，构建模态平衡约束与混合注意力路径，有效融合不同模态特征，增强模态间的协同表达能力，提升图文与视频风险内容识别的整体性能。

**3.** **基于个性化调制器的LoRA大模型微调方法**：为应对数据孤岛和隐私保护等问题，项目构建了联邦式的双通道多模态建模框架FEDPM，结合知识蒸馏与个性化学习策略，显著提升模型在分布异构场景下的跨域泛化能力，实现“数据不出域，模型共进化”的协同训练新范式。

结合BLIP2的轻量级参数微调、基于CLIP的伪造图像构建、三重结构相似性对齐等关键方法，项目在模型构建上兼顾性能与计算效率，形成了统一架构、多模态融合、可扩展性强的技术体系，具有良好的落地潜力与推广前景。

## 应用推广

本项目成果在多模态风险内容检测领域具有广泛的实际应用价值。系统可应用于社交平台虚假新闻识别、诈骗电话与短信检测、伪造音视频监测等多个重点场景，有效提升内容审核、平台治理和舆情管理能力，服务于网络空间治理、安全监管与数字社会建设。

系统架构支持模块化部署和联邦协同更新，可按需适配不同终端与场景，具备从云端平台到边缘设备的灵活应用能力。未来可广泛推广至媒体平台、公安监管、网络安全、舆情监测等领域，为政企机构提供智能化的多模态风险识别解决方案。

## 作品展望

随着多模态生成技术与大模型能力的不断增强，虚假信息的伪造能力持续提升，内容形式更具迷惑性与传播性。本项目在现有成果基础上，后续将在高水平期刊上**发表论文3篇**，**提交发明专利2项**。

同时，我们将探索大模型在低资源场景下的轻量化部署策略与差异化更新机制，推动模型向移动端、边缘设备等轻型终端迁移，助力构建高效、可控、安全的智能化内容识别体系。依托团队现有的研究基础与技术成果，项目将在更大范围内推动多模态大模型技术在实际场景中的应用与落地，为建设清朗健康的数字环境提供有力支撑。

# 参考文献

1. Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, “Eann: Event adversarial neural networks for multi-modal fake news detection,” in ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2018, pp. 849–857.
2. S. Singhal, R.R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, “Spotfake: A multi-modal framework for fake news detection,” in International Conference on Multimedia Big Data, 2019, pp. 39–47.
3. T. Zhang, D. Wang, H.Chen, Z.Zeng, W.Guo, C.Miao, and L. Cui, “Bdann: Bert-based domain adaptation neu ral network for multi-modal fake news detection,” in In ternational Joint Conference on Neural Networks, 2020, pp. 1–8.
4. D. Khattar, J.S. Goud, M. Gupta, and V. Varma, “Mvae: Multimodal variational autoencoder for fake news de tection,” in The world Wide Web Conference, 2019, pp. 2915–2921.
5. P. Wei, F. Wu, Y. Sun, H. Zhou, and X. Jing, “Modal ity and event adversarial networks for multi-modal fake news detection,” IEEE Signal Processing Letters, vol. 29, pp. 1382–1386, 2022.
6. Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, “Cross-modal ambiguity learning for multi modal fake news detection,” in ACM Web Conference, 2022, pp. 2897–2905.
7. L. Peng, S. Jian, D. Li, and S. Shen, “Mrml: Multimodal rumor detection by deep metric learning,” in IEEE In ternational Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.
8. X. Zhou, J. Wu, and R. Zafarani, “Safe: similarity aware multi-modal fake news detection (2020),” Preprint. arXiv, vol. 200304981, 2020.
9. J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, “De tecting fake news by exploring the consistency of mul timodal data,” Information Processing & Management, vol. 58, no. 5, pp. 102610, 2021.
10. Y. Shao, J. Sun, T. Zhang, Y. Jiang, J. Ma, and J. Li, “Fake news detection based on multi-modal classifier ensemble,” in International Workshop on Multimedia AI against Disinformation, 2022, pp. 78–86.
11. X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, “Farewell to mutual information: Variational distillation for cross-modal person re-identification,” in IEEE/CVF Conference on Computer Vision and Pattern Recogni tion, 2021, pp. 1522–1531.
12. X. Tian, Z. Zhang, C. Wang, W. Zhang, Y. Qu, L. Ma, Z. Wu, Y. Xie, and D. Tao, “Variational distillation for multi-view learning,” arXiv preprint arXiv:2206.09548, 2022.
13. W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, “Evidence aware fake news detection with graph neural networks,” in ACM WebConference, 2022, pp. 2501–2510.
14. J. Chen and G. Kou, “Attribute and structure preserv ing graph contrastive learning,” in AAAI Conference on Artificial Intelligence, 2023, number 6, pp. 7024–7032.
15. C. Du, T. Li, Y. Liu, Z. Wen, T. Hua, Y. Wang, H. Zhao, Improving multi-modal

learning with uni-modal teachers, 2021, arXiv preprint arXiv:2106.11059.

1. X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via

on-the-fly gradient modulation, in: IEEE/CVF Conference on Computer Vision

and Pattern Recognition, 2022, pp. 8238–8247.

1. Y. Wei, R. Feng, Z. Wang, D. Hu, Enhancing multimodal cooperation via sample level modality valuation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27338–27347.
2. Iacob, A., Sani, L., Kurmanji, M., Shen, W. F., Qiu, X., Cai, D., Gao, Y., and Lane, N. D. DEPT: Decoupled Embed dings for Pre-training Language Models. In International Conference on Learning Representations, 2025.
3. Weller, O., Marone, M., Braverman, V., Lawrie, D., and Van Durme, B. Pretrained Models for Multilingual Feder ated Learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Com putational Linguistics: Human Language Technologies, 2022.
4. Ye, R., Ge, R., Zhu, X., Chai, J., Du, Y., Liu, Y., Wang, Y., and Chen, S. FedLLM-Bench: Realistic Benchmarks for Federated Learning of Large Language Models. In Ad vances in Neural Information Processing Systems, 2024.
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm´ an, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual repre sentation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
6. Wang, Z., Lipton, Z. C., and Tsvetkov, Y. On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020b.
7. Jiang, Y., Koneˇ cn` y, J., Rush, K., and Kannan, S. Improving Federated Learning Personalization via Model Agnostic Meta Learning. arXiv preprint arXiv:1909.12488, 2019.
8. Lee, R., Kim, M., Li, D., Qiu, X., Hospedales, T., Husz´ar, F., and Lane, N. FedL2P: Federated Learning to Personal ize. Advances in Neural Information Processing Systems, 2023.
9. Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, Xiao-Yuan Jing:Balanced Multi-modal Learning with Hierarchical Fusion for Fake News Detection. Pattern Recognit. 164: 111485 (2025)
10. X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via on-the-fly gradient modulation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.
11. Y. Wei, R. Feng, Z. Wang, D. Hu, Enhancing multimodal cooperation via samplelevel modality valuation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27338–27347.
12. R. Xu, R. Feng, S.-X. Zhang, D. Hu, Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.
13. Lee, R., Kim, M., Rezk, F., Li, R., Venieris, S. I., & Hospedales, T. (2025). FedP²EFT: Federated learning to personalize parameter efficient fine-tuning for multilingual LLMs. arXiv.
14. Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, Xiao-Yuan Jing:

Balanced Multi-modal Learning with Hierarchical Fusion for Fake News Detection. Pattern Recognit. 164: 111485 (2025)