FedAMB: Adaptive Modality Balancing for Dominance-Robust Multimodal Federated Distillation

Seungjin Han[1] , **Juyeob Lee**[1] , **Sangmin Lee**[2,*] , **Eunil Park**[1,3,*]
[1]Department of Applied Artificial Intelligence, Sungkyunkwan University, Korea
[2]Department of Computer Science and Engineering, Korea University, Korea
[3]Department of Computer Science and Engineering, Jaume I University, Spain
seungjinh99@g.skku.edu, juyblee@gmail.com, sangmin-lee@korea.ac.kr, eunilpark@skku.edu

Abstract

Federated knowledge distillation (Fed-KD) exchanges distilled predictions on a shared proxy dataset, often reducing communication and accommodating heterogeneous client architectures. However, in multimodal federated learning (MFL), *modality dominance*, where models over-rely on an easier modality, can bias local optimization and *contaminate* the aggregated global teacher, degrading both multimodal accuracy and robustness to missing modalities, especially under non-IID heterogeneity. We propose *Federated Adaptive-Modality Balancing* (FedAMB). At the client side, *Selective-Modality Regulation* (SMR) models dominance as a state-dependent phenomenon and intervenes only when it destabilizes training, strengthening weak modalities without over-regularization. At the server side, *Component-wise Modality Distillation* (CMD) regulates how aggregated knowledge is transferred to each modality branch, preventing the propagation of fusion-biased teachers while directly improving unimodal representations. Experiments on CREMA-D, AVE, and UR-FUNNY show that FedAMB improves multimodal accuracy and missing-modality robustness.

1 Introduction

The expanding application of artificial intelligence has increased the demand for robust data protection mechanisms. Federated Learning (FL) has emerged as a pivotal solution to this challenge, particularly in privacy-sensitive domains such as healthcare [Xu *et al.*, 2021], finance [Long *et al.*, 2020], and personalized services [Mammen, 2021]. In a typical FL setting, a global model is distributed from a central server to multiple clients [McMahan *et al.*, 2017], where it is trained locally using the private data of the clients. Instead of sharing sensitive raw data, clients transmit model updates, such as weights [Kairouz *et al.*, 2021] or gradients [Wen *et al.*, 2017], back to the server. The server aggregates these updates to refine the global model and initiates subsequent training rounds. This collaborative paradigm enables multiple organizations to build robust models from diverse data sources without exposing confidential information [Rieke *et al.*, 2020]. Despite their success, conventional FL approaches that rely on model-update exchanges suffer from several limitations. Transmitting the full model parameters of large-scale networks incurs a substantial communication overhead. Moreover, enforcing a homogeneous model architecture across all clients restricts flexibility in the presence of heterogeneous data distributions and computational resources [Mora *et al.*, 2022]. The direct sharing of model updates also raises concerns regarding information leakage and intellectual property protection [Yin *et al.*, 2021]. To alleviate these issues, Federated Knowledge Distillation (Fed-KD) has been proposed as a promising alternative [Li *et al.*, 2024]. Fed-KD exchanges condensed knowledge, such as intermediate features or output logits, rather than full model parameters, thereby reducing the communication costs and mitigating privacy risks.

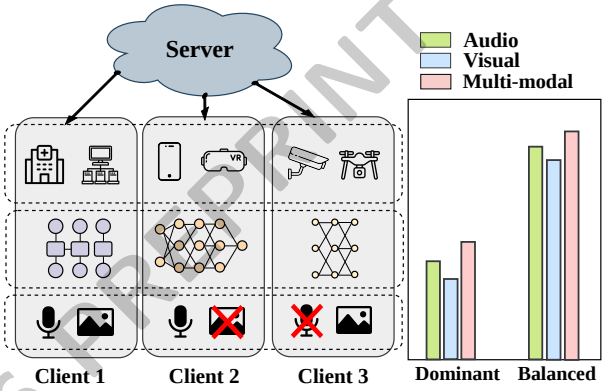


Figure 1: Illustrations of Federated learning and key challenges. Left: Heterogeneity in model architectures and modality availability, including missing modalities. Right: Performance degradation caused by modality dominance.

Meanwhile, in real-world FL applications, data are often inherently multimodal and comprise multiple information sources, such as images, text, and speech. Learning from multiple modalities enables models to capture complementary information and achieve superior performance compared with unimodal approaches. Motivated by this, Multimodal Federated Learning (MFL) has emerged as a promising

*Corresponding authors

research direction that combines multimodal representation learning with the privacy-preserving advantages of FL [Che *et al.*, 2023]. However, the multimodal–federated optimization interaction introduces challenges that do not occur in unimodal FL. A primary issue is *modality dominance*, where a single strong or easier-to-learn modality disproportionately governs the learning process, suppressing complementary signals from other modalities and leading to suboptimal multimodal fusion [Wang *et al.*, 2020] (Fig. 1, right). In practical terms, a client model may appear to perform well by relying mostly on its locally dominant modality, while the remaining modalities are under-trained and become unreliable under missing-modality inference. This problem is significantly exacerbated in federated environments with non-independent and identically distributed (Non-IID) data [Zhao *et al.*, 2018], where different clients may develop different modality-specific biases. Consequently, in Fed-KD-based MFL, these client-specific biases can be propagated and amplified through server-side aggregation, contaminating the global teacher and ultimately degrading the performance and robustness of the global model [Ouyang *et al.*, 2023].

Another critical challenge arises from missing modalities during deployment. Owing to sensor failures or resource constraints, one or more modalities may be unavailable at inference time [Wu *et al.*, 2024]. As a result, clients exhibit inconsistent modality configurations (Figure 1, left). Therefore, a practical MFL framework must deliver reliable performance even under unimodal inference without incurring communication overheads [Wang *et al.*, 2024]. Prior studies have attempted to mitigate these challenges; however, existing methods often lack an aggregation-aware perspective. In particular, they overlook how dominance-induced client bias can distort server aggregation and contaminate the global teacher, thereby propagating biased knowledge back to clients. Moreover, prior dominance mitigation typically assumes a static imbalance and applies uniform interventions (e.g., fixed weighting or modality dropping), whereas dominance in MFL is inherently state-dependent across clients and rounds. As a result, indiscriminate regulation may unnecessarily disrupt stable multimodal fusion.

To address these challenges, we propose a novel MFL framework built upon the Fed-KD paradigm, called *Federated Adaptive Modality Balancing* (FedAMB). At the client level, *Selective-Modality Regulation* (SMR) mitigates dominance at its source by selectively regulating biased local learning dynamics. At the global stage, *Component-wise Modality Distillation* (CMD) decomposes and regulates the flow of global knowledge, preventing client-level fusion bias from being propagated through the global teacher and directly strengthening unimodal representations. Through the joint application of SMR and CMD, FedAMB enhances the expressiveness of unimodal representations while stabilizing multimodal fusion under heterogeneous and incomplete modality settings. Extensive experiments on the CREMA-D, AVE, and UR-FUNNY datasets, conducted under varying degrees of data heterogeneity, demonstrate that FedAMB consistently outperforms existing methods in terms of multimodal accuracy and unimodal robustness. Our contributions are summarized as follows:

- **First contamination-aware view of modality dominance in Fed-KD MFL:** To the best of our knowledge, this is the first work to cast modality dominance as global-teacher contamination in Fed-KD-based MFL, showing that dominance-induced client bias is amplified through server aggregation and degrades both multimodal accuracy and unimodal robustness.
- **FedAMB: stage-wise dominance mitigation with selective intervention and branch-wise distillation:** We propose FedAMB, which (i) activates SMR only when dominance destabilizes local learning and (ii) introduces CMD to distill global knowledge into each modality branch, preventing fusion-biased teacher signals from suppressing unimodal representations.
- **Robust improvements under realistic FL conditions:** Across CREMA-D, AVE, and UR-FUNNY, FedAMB consistently improves multimodal performance and unimodal robustness under heterogeneous Non-IID partitions and missing-modality settings, outperforming strong baselines by a stable margin.

2 Related Work

2.1 Federated Learning Paradigms

FL is a privacy-enhancing technique that enables collaborative model training on decentralized data [Li *et al.*, 2020a]. The most common approach involves clients sharing parameter updates, such as weights [Karimireddy *et al.*, 2020] or gradients [Yuan and Ma, 2020] from their locally trained models with a server to build a single global model. These conventional FL approaches, which are based on exchanging global models, suffer from several inherent weaknesses [Li *et al.*, 2020b]. First, the increasing size of deep neural networks leads to prohibitive communication overheads when model weights are directly exchanged [Lin *et al.*, 2018; Alistarh *et al.*, 2017]. Second, it imposes a rigid requirement for a homogeneous model architecture across all participating clients [Diao *et al.*, 2021]. Finally, the direct exchange of model updates creates significant security vulnerabilities, exposing the system to threats such as model inversion attacks, which can potentially reconstruct private training data [Fraboni *et al.*, 2021]. Communication, architectural flexibility, and security challenges are key factors that impede the practical deployment of FL.

2.2 Knowledge Distillation with Proxy Dataset

To overcome the limitations of update-sharing FL, Fed-KD was proposed as an effective alternative [Qin *et al.*, 2024]. The core idea of Fed-KD is to exchange "knowledge," such as model outputs (logits) or features [Yang *et al.*, 2023] from intermediate layers, instead of heavy model parameters [Lin *et al.*, 2020]. A shared public dataset is necessary for effectively exchanging knowledge among clients who cannot share their private data with others. Therefore, a concept of a proxy dataset was proposed that all participants can access [Hinton *et al.*, 2015a]. This dataset enables clients to generate predictions on identical inputs, which can then be aggregated by the server to form global teacher knowledge for distillation [Hinton *et al.*, 2015b].

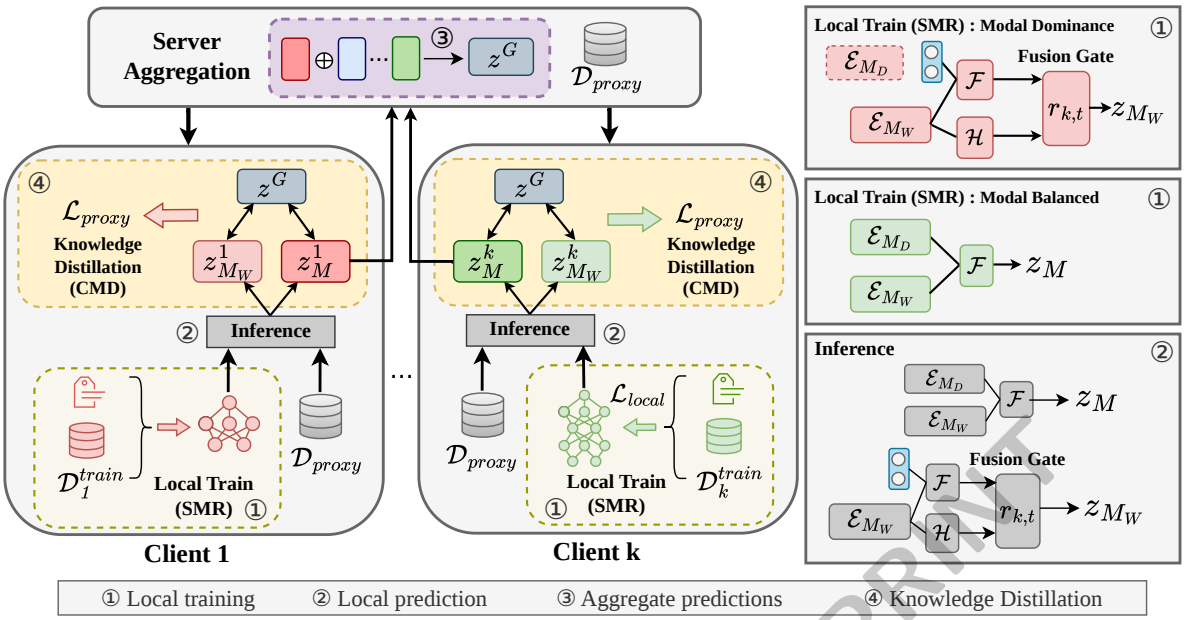


Figure 2: Overview of the FedAMB architecture. The numbered steps indicate the communication rounds. (1) Each client identifies modality dominance and performs local training using Selective-Modality Regulation. (2) Client-generated multimodal logits from the proxy dataset are transmitted to the server. (3) The server aggregates these logits to create the global knowledge and distributes the results. (4) Clients distill the global knowledge using the Component-wise Modality Distillation.

2.3 Multimodal Federated Learning

The recent explosion of data from the Internet, sensors, and mobile devices has significantly increased the importance of multimodal learning, which is a field dedicated to leveraging diverse data types in unison [Baltrušaitis *et al.*, 2018]. MFL has arisen from this background as a novel research field that integrates the privacy advantages of FL with the powerful analytical capabilities of multimodal learning [Yu *et al.*, 2023]. However, MFL faces a set of unique and multifaceted problems that are distinct from those in unimodal FL [Adam *et al.*, 2025]. One of these is modality heterogeneity, in which clients possess different sets of modalities, leading to architectural discrepancies that hinder the aggregation of models. Furthermore, MFL often suffers from modality dominance, a phenomenon in which a single dominant modality can overwhelm the learning process and suppress valuable signals from other modalities. Several studies have attempted to address these challenges in MFL. One line of research focuses on handling incomplete or missing modalities. For instance, some approaches utilize contrastive learning to learn robust multimodal representations even when certain modalities are unavailable [Che *et al.*, 2024], while others investigate methods for generating or restoring missing-modality data at the client level for training [Xiong *et al.*, 2023]. Another line of research addresses the modality heterogeneity and imbalance. One study proposed the use of submodular functions to promote modality diversity during client selection and introduced a modal enhancement loss to facilitate the learning of weak modalities [Fan *et al.*, 2024]. However, this approach is based on conventional FL, which exchanges model updates.Existing Fed-KD methods can falter in MFL because dominance-biased client predictions may be aggregated into a fusion-biased global teacher, which then suppresses weak-modality learning again during distillation [Sarfraz *et al.*, 2021]. Different from parameter-aggregation-based MFL methods, FedAMB follows a Fed-KD protocol that aggregates proxy predictions as the primary communication signal, while exchanging lightweight class-wise prototypes for dominance identification. With an aggregation-aware stage-wise design (SMR+CMD), FedAMB mitigates this specific teacher-contamination failure mode and improves both multimodal performance and unimodal robustness.

3 Methodology

3.1 Problem Formulation

We consider a MFL system with K clients and a modality set $\mathcal{M}$. Client k holds a private labeled dataset

$$\mathcal{D}_k = \{(\{x_{m,i}^k\}_{m\in\mathcal{M}}, y_i^k)\}_{i=1}^{N_k}, \tag{1}$$

where N_k is the number of samples, and $y_i^k \in \{1,\dots,N_c\}$ is the class label. We split $\mathcal{D}_k$ into a training set $\mathcal{D}_k^{train}$ and a test set $\mathcal{D}_k^{test}$. Following the federated knowledge distillation (Fed-KD) paradigm, we assume an unlabeled proxy dataset that is accessible to all clients:

$$\mathcal{D}_{proxy} = \{\{x_{m,i}^{proxy}\}_{m\in\mathcal{M}}\}_{i=1}^{N_{proxy}}. \tag{2}$$

The goal is to collaboratively improve each client model $\mathcal{C}_{\theta_k}$ without sharing raw private data, such that the resulting model achieves (i) high multimodal accuracy under the available modalities and (ii) robust unimodal performance when some modalities are absent at inference.

Algorithm 1 FedAMB (Per Round)

Input: total rounds T, dominance threshold γ, fusion-instability threshold τ, KD temperature T_{kd}
Output: client models $\{\theta_k\}$

Server maintains global teacher Z^G and global prototypes $\{P^G\}$
for $t = 1$ **to** T **do**
  Server samples clients $\mathcal{S}_t$ and broadcasts $(Z^G, \{P^G\})$
  for each client $k \in \mathcal{S}_t$ **do**
    Compute $PAcc_{k,t}(m)$ using a private sample and $\{P^G\}$
    $M_D^{k,t} \leftarrow \arg\max_m PAcc_{k,t}(m)$, $M_W^{k,t} \leftarrow \arg\min_m PAcc_{k,t}(m)$
    $s_{k,t} \leftarrow [\, PAcc_{k,t}(M_D^{k,t}) > \gamma \cdot PAcc_{k,t}(M_W^{k,t})\,]$
    Run local training with SMR if $s_{k,t} = 1$; else standard multimodal training
    Upload proxy logits z^k_{proxy} and prototype stats to server
  end for
  Server aggregates $\{z^k_{proxy}\}$ to update Z^G and aggregates prototypes to update $\{P^G\}$
  Server broadcasts updated Z^G
  for each client $k \in \mathcal{S}_t$ **do**
    Compute proxy logits (z^k_M, z^k_W)
    $\mathcal{L}^k_{\text{CMD}} = \sum_{u\in\{M,W\}} KL(z^k_u, Z^G)$
    Update θ_k by minimizing $\mathcal{L}^k_{CMD}$
  end for
end for
return $\{\theta_k\}$

3.2 FedAMB: Overall Framework

FedAMB is a stage-wise Fed-KD framework that mitigates dominance-induced client bias before aggregation and controls the propagation of global knowledge back to each modality branch. Communication round t consists of the following steps: Figure 2 provides an overview of the pipeline, and Algorithm 1 summarizes the detailed round procedure. Specifically, (0) *Prototype Upload and Aggregation*: each selected client computes modality- and class-specific local prototypes, uploads $(P^{k,c}_{m,t}, N^{k,c})$ to the server, and the server aggregates them into global prototypes $\{P^{G,c}_{m,t}\}$ and broadcasts them back for prototype-based dominance identification; (1) *Selective-Modality Regulation (SMR)*: using $\{P^{G,c}_{m,t}\}$, client k estimates the dominance state $s_{k,t}$ and, if dominated, suppresses the dominant modality by training with a weak-modality logit (masking the dominant modality); (2) *Proxy Logit Upload*: clients compute multimodal logits on the proxy samples and upload them as distilled knowledge; (3) *Teacher Construction and Distribution*: the server aggregates uploaded logits to construct the global teacher z^G_t and broadcasts it to clients; and (4) *Component-wise Modality Distillation (CMD)*: each client distills the global teacher into both its multimodal logits and weak unimodal logits, strengthening the weak branch while preventing fusion-induced bias from being propagated through global aggregation.

3.3 Dominance Identification as a State Variable

Let $m \in \mathcal{M}$ denote modality. Client k maintains a modality-specific encoder $\mathcal{E}^k_m$ that extracts features $f^k_{m,i,t} = \mathcal{E}^k_m(x^k_{m,i})$ at round t. For each class c, client k computes the local prototype as

$$P^{k,c}_{m,t} = \frac{1}{N^{k,c}} \sum_{i:y^k_i=c} f^k_{m,i,t} \tag{3}$$

where $N^{k,c}$ is the number of samples of class c at client k (computed on $\mathcal{D}^{train}_k$). Clients upload $\{(P^{k,c}_{m,t}, N^{k,c})\}$ to the server, which forms a class-weighted global prototype:

$$P^{G,c}_{m,t} = \frac{\sum_{k\in\mathcal{K}^c} N^{k,c} P^{k,c}_{m,t}}{\sum_{k\in\mathcal{K}^c} N^{k,c}}, \quad \mathcal{K}^c = \{k \mid N^{k,c} > 0\}. \tag{4}$$

Prototype-based accuracy. Given global prototypes $\{P^{G,c}_{m,t}\}_c$, client k assigns a prototype label using cosine similarity as follows:

$$\hat{y}^k_{m,i,t} = \arg\max_{c\in\{1,\dots,N_c\}} \frac{f^k_{m,i,t} \cdot P^{G,c}_{m,t}}{\|f^k_{m,i,t}\|_2 \, \|P^{G,c}_{m,t}\|_2}. \tag{5}$$

To avoid ambiguity and keep the computation lightweight, we estimate the dominance metric on a sampled labeled mini-batch $\mathcal{B}_{k,t} \subset \mathcal{D}^{train}_k$:

$$PAcc_{k,t}(m) = \frac{1}{|\mathcal{B}_{k,t}|} \sum_{(x^k_i, y^k_i)\in\mathcal{B}_{k,t}} \mathbb{I}\left(\hat{y}^k_{m,i,t} = y^k_i\right). \tag{6}$$

Dominance state. We define the dominant and weak modalities for client k at round t as

$$\begin{aligned} M^{k,t}_D &= \arg\max_{m\in\mathcal{M}} PAcc_{k,t}(m), \\ M^{k,t}_W &= \arg\min_{m\in\mathcal{M}} PAcc_{k,t}(m). \end{aligned} \tag{7}$$

and the dominance indicator

$$s_{k,t} = \mathbb{I}\Big(PAcc_{k,t}(M^{k,t}_D) > \gamma\, PAcc_{k,t}(M^{k,t}_W)\Big), \tag{8}$$

where $\gamma > 1$ is the threshold.

3.4 Selective-Modality Regulation (SMR)

Let $\mathcal{F}^k$ be the fusion head producing multimodal logits from modality features, and let $\mathcal{H}^k_m$ be an auxiliary unimodal head for modality m. At round t, for a sample i, the multimodal logit is

$$z^k_{M,i,t} = \mathcal{F}^k\left(\{f^k_{m,i,t}\}_{m\in\mathcal{M}}\right). \tag{9}$$

When $s_{k,t} = 0$ (non-dominant state), the client performs standard multimodal training with $z^k_{M,i,t}$.

When $s_{k,t} = 1$ (dominant state), SMR suppresses the dominant modality by masking its features and constructing an alternative training logit for the weak modality. We define masked features

$$\tilde{f}^k_{m,i,t} = \begin{cases} \mathbf{0}, & \text{if } m = M^{k,t}_D, \\ f^k_{m,i,t}, & \text{otherwise,} \end{cases} \tag{10}$$

and compute a fusion logit without the dominant modality as follows:

$$\tilde{z}^k_{M,i,t} = \mathcal{F}^k\left(\{\tilde{f}^k_{m,i,t}\}_{m\in\mathcal{M}}\right). \tag{11}$$

We also compute the weak-modality unimodal logit:

$$z^k_{H,i,t} = \mathcal{H}^k_{M^{k,t}_W}\left(f^k_{M^{k,t}_W,i,t}\right). \tag{12}$$

Let $\mathrm{Acc}^{k,t}_M$ and $\mathrm{Acc}^{k,t}_{MD}$ denote the training accuracy of the multimodal prediction and the dominant-modality prediction, respectively. To avoid over-intervention within the dominant states, we activate a fusion-instability gate only when multimodal fusion becomes less reliable than the dominant-modality prediction:

$$r_{k,t} = \mathbb{I}\left(\frac{\mathrm{Acc}^{k,t}_{MD}}{\mathrm{Acc}^{k,t}_M} > \tau\right). \tag{13}$$

Then, we incorporate the masked-fusion signal $\tilde{z}^k_{M,i,t}$ as a corrective term and construct the weak-modality training logit in a single form:

$$z^k_{W,i,t} = z^k_{H,i,t} + r_{k,t}\,\tilde{z}^k_{M,i,t}. \tag{14}$$

We choose the local training logit as

$$z^k_{L,i,t} = (1 - s_{k,t})\,z^k_{M,i,t} + s_{k,t}\,z^k_{W,i,t}, \tag{15}$$

and minimize the local cross-entropy loss

$$\mathcal{L}^{k,t}_{local} = \mathcal{L}_{CE}(z^k_{L,i,t}, y^k_i). \tag{16}$$

3.5 Component-wise Modality Distillation (CMD)

After the SMR-based local updates, the clients participate in Fed-KD by computing the proxy logits. Client k uploads its multimodal proxy logits $\{z^k_{M,t}(x)\}_{x\in\mathcal{D}_{proxy}}$. The server constructs a global teacher by averaging the logits as follows:

$$z^G_t(x) = \frac{1}{|\mathcal{S}_t|}\sum_{k\in\mathcal{S}_t} z^k_{M,t}(x), \tag{17}$$

where $\mathcal{S}_t$ is the set of participating clients in round t.

The global teacher serves as a soft ensemble target capturing class structures and unbiased boundaries, rather than requiring exact reproduction of multimodal data. Thus, distillation regularizes unimodal representations toward this structure instead of transferring full multimodal information. This allows CMD to stabilize weak-modality learning without assuming teacher-student capacity alignment. Each client then performs distillation at a fixed temperature $T_{\mathrm{kd}} > 0$:

$$\mathcal{L}^{k,t}_{multi\text{-}kd} = T^2_{\mathrm{kd}}\,D_{KL}\left(\sigma\left(\frac{z^G_t}{T_{\mathrm{kd}}}\right)\,\middle\|\,\sigma\left(\frac{z^k_{M,t}}{T_{\mathrm{kd}}}\right)\right), \tag{18}$$

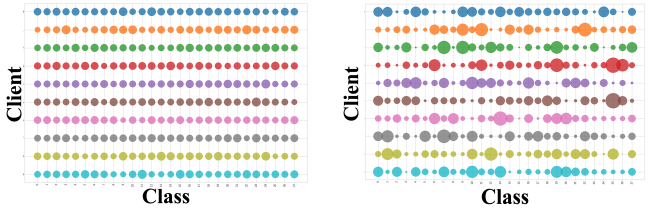
Additionally, the same teacher is distilled into the weak-modality branch:

$$\mathcal{L}^{k,t}_{uni\text{-}kd} = T^2_{\mathrm{kd}}\,D_{KL}\left(\sigma\left(\frac{z^G_t}{T_{\mathrm{kd}}}\right)\,\middle\|\,\sigma\left(\frac{z^k_{W,t}}{T_{\mathrm{kd}}}\right)\right). \tag{19}$$

The proxy loss is

$$\mathcal{L}^{k,t}_{proxy} = \mathcal{L}^{k,t}_{multi\text{-}kd} + \mathcal{L}^{k,t}_{uni\text{-}kd}. \tag{20}$$

The CMD explicitly regulates the transfer of global knowledge to each modality component, strengthening unimodal representations while avoiding the blind propagation of fusion bias through the global teacher.



(a) IID (b) Non-IID

Figure 3: Illustrations of the class distribution of each client in (a) IID and (b) Non-IID settings.

4 Experiments

Datasets We evaluate FedAMB on two **audio–visual** datasets and one **audio–visual–text** dataset. For audio–visual learning, we use CREMA-D [Cao *et al.*, 2014] (emotion recognition) and AVE [Tian *et al.*, 2018] (event detection), where modality dominance is observed, thus providing a suitable testbed for our dominance-aware MFL setting. To further assess scalability beyond two modalities, we additionally use UR-FUNNY, an **audio–visual–text** benchmark, to evaluate FedAMB under N-modal learning.

Implementation Details We model client-level statistical heterogeneity using Dirichlet partitioning over class labels and define the non-IID setting with $\alpha = 1.0$. Unless otherwise specified, we run federated learning with 10 clients for 100 communication rounds, where 4 clients are randomly selected per round, and set the dominance threshold to $\gamma = 1.5$. To simulate an in-domain FL environment under the Fed-KD protocol, we sample 20% of each dataset as a public, unlabeled proxy set for knowledge exchange, and distribute the remaining 80% across clients as private local datasets.

Baselines To evaluate, we include FedMD [Li and Wang, 2019], FedProto [Tan *et al.*, 2022], FedIoT-KD [Zhang *et al.*, 2021], and DaFKD [Wang *et al.*, 2023] as representative Fed-KD and prototype-based baselines for MFL. In addition, we compare with Random Modal Drop (RMD) [Alfasly *et al.*, 2022] and FedPMR [Fan *et al.*, 2023] as modality-dominance mitigation baselines.

4.1 Evaluation Results

Table 1 reports the performance comparison of CREMA-D, AVE, and UR-FUNNY under IID and non-IID settings. Overall, FedAMB consistently outperforms strong baselines across all datasets, demonstrating robust gains under both homogeneous and heterogeneous client distributions. In particular, FedAMB improves over the best baseline by +5.29/+1.64 points on CREMA-D (IID/Non-IID), +0.30/+0.15 points on AVE, and +0.13/+0.56 points on UR-FUNNY, respectively. These results indicate that FedAMB generalizes beyond audio–visual settings to N-modality scenarios and yields stable improvements under federated heterogeneity. Unlike RMD and FedPMR, FedAMB explicitly suppresses dominance-induced bias in aggregated logit knowledge via SMR and CMD, leading to a more stable and robust performance under heterogeneity. To further analyze

Method	CREMA-D		AVE		UR-FUNNY	
	IID	Non-IID	IID	Non-IID	IID	Non-IID
FedMD	41.08 ± 0.07	39.17 ± 0.14	36.11 ± 0.09	35.00 ± 0.11	58.14 ± 0.02	53.22 ± 0.10
FedProto	38.10 ± 0.04	36.76 ± 0.16	34.08 ± 0.08	30.37 ± 0.19	55.14 ± 0.06	52.86 ± 0.07
FedIoT	36.29 ± 0.13	33.41 ± 0.14	32.21 ± 0.06	31.82 ± 0.15	54.87 ± 0.04	52.73 ± 0.03
RMD	40.26 ± 0.05	40.55 ± 0.17	36.44 ± 0.03	31.14 ± 0.14	56.59 ± 0.10	53.37 ± 0.11
FedPMR	39.03 ± 0.21	37.14 ± 0.28	20.26 ± 0.33	19.28 ± 0.38	57.48 ± 0.08	54.10 ± 0.09
DaFKD	44.82 ± 0.08	41.09 ± 0.12	35.91 ± 0.05	33.84 ± 0.09	57.96 ± 0.06	54.62 ± 0.11
FedAMB (Ours)	$\mathbf{50.11 \pm 0.12}$	$\mathbf{42.73 \pm 0.16}$	$\mathbf{36.74 \pm 0.11}$	$\mathbf{35.15 \pm 0.14}$	$\mathbf{58.27 \pm 0.14}$	$\mathbf{55.18 \pm 0.10}$

Table 1: Performance comparison (%) on CREMA-D, AVE, and UR-FUNNY under IID and Non-IID settings. Results are reported as mean ± standard deviation over 3 random seeds.

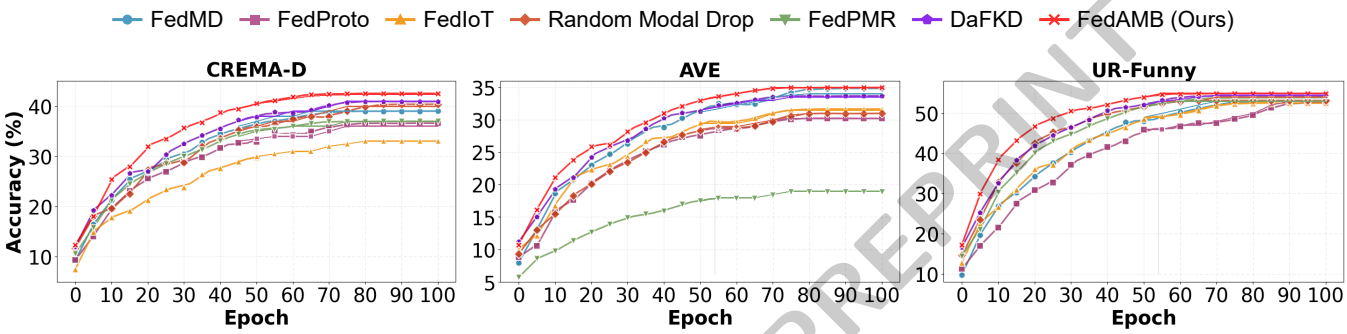


Figure 4: Average test accuracy comparison of FedMD, FedProto, FedIoT, RMD, FedPMR, DaFKD, and FedAMB on three datasets under the Non-IID setting.

Setting	Method	Audio	Visual
IID	FedMD	40.07±0.13	15.03±0.03
	FedProto	35.46±0.08	16.53±0.01
	FedIoT	32.13±0.16	16.69±0.05
	RMD	38.56±0.10	22.65±0.08
	FedPMR	37.61±0.18	20.20±0.07
	DaFKD	40.29±0.11	13.95±0.02
	FedAMB (Ours)	**40.78** ±0.06	**34.12**±0.05
Non-IID	FedMD	38.15±0.15	15.17±0.05
	FedProto	34.73±0.12	16.13±0.08
	FedIoT	32.97±0.16	15.59±0.04
	RMD	33.55±0.13	25.08±0.12
	FedPMR	37.11±0.14	18.73±0.11
	DaFKD	38.90±0.15	15.85±0.04
	FedAMB (Ours)	**39.52**±0.13	**27.33**±0.15

Table 2: Unimodal performance comparison (%) on CREMA-D dataset in IID and Non-IID settings.

the training dynamics, we visualize the convergence behavior of all methods for each dataset. Figure 4 presents the test accuracy curves for the communication rounds. Notably, although FedAMB employs a state-dependent loss, it remains stable throughout the training and consistently converges under federated heterogeneity.

4.2 Unimodal Performance Analysis

To further investigate the effectiveness of FedAMB, we conduct a unimodal performance analysis on the CREMA-D dataset. This analysis evaluates the model's accuracy on individual audio and visual modalities by replacing the feature vector from the other modality with zero vectors. Table 2 presents the accuracy of unimodal inferences. In the IID setting, FedAMB improves the visual modality by more than +11.47 points and achieves the highest audio accuracy in both IID (40.78%) and non-IID (39.52%) settings. The consistent top performance across both modalities, especially the improvement in the weak modality (visual), underscores the capability of FedAMB to effectively leverage and balance information from different sources. This holds for both IID and the more challenging non-IID settings, further validating the robustness and efficacy of the FedAMB framework.

4.3 Robustness under a Public Proxy Dataset

To assess the practicality of FedAMB when an in-domain proxy is unavailable, we consider a domain-mismatched public proxy scenario where all methods distill knowledge from CMU-MOSEI [Zadeh *et al.*, 2018] (Table 3). Despite the distribution shift between the proxy and private client data, FedAMB remains consistently competitive and achieves the best performance across both datasets and client distributions. Specifically, FedAMB improves over the strongest baseline by +4.74/+0.71 points on CREMA-D (IID/Non-IID) and by +0.93/+0.97 points on AVE. This suggests that FedAMB is

Method	CREMA-D		AVE	
	IID	Non-IID	IID	Non-IID
FedMD	36.49	34.24	32.67	31.12
FedProto	36.85	33.76	29.03	25.46
FedIoT	33.17	28.49	26.84	26.79
RMD	37.84	37.11	33.28	28.31
FedPMR	35.33	33.73	18.12	16.80
DaFKD	42.98	38.24	31.65	29.49
FedAMB (Ours)	**47.72**	**38.95**	**34.21**	**32.09**

Table 3: Performance comparison (%) on CREMA-D and AVE under IID and Non-IID settings. All methods are evaluated using the same public proxy dataset, CMU-MOSEI.

Dataset	SMR	CMD	Multi	Audio	Visual
CREMA-D	×	×	41.08	40.07	15.03
	✓	×	44.26	40.01	25.59
	×	✓	41.20	40.35	15.07
	✓	✓	**50.11**	**40.78**	**34.12**
AVE	×	×	36.11	29.20	10.33
	✓	×	35.33	27.53	13.41
	×	✓	36.09	29.50	10.62
	✓	✓	**36.74**	**29.88**	**13.92**

Table 4: Multimodal and unimodal accuracy (%) of ablation study on FedAMB components in the CREMA-D and AVE datasets: Selective-Modality Regulation (SMR) and Component-wise Modality Distillation (CMD).

Data	CREMA-D	AVE	UR-FUNNY
w/o Dual-Head	50.02	32.29	56.47
w/ Dual-Head	**50.10**	**36.74**	**58.27**

Table 5: Multimodal accuracy (%) of the Dual-Head ablation on CREMA-D, AVE, and UR-FUNNY.

robust to proxy mismatches and can still suppress dominance-induced bias even when the global teacher is constructed from an external public dataset, supporting its deployment in realistic Fed-KD settings.

4.4 Ablation Study

Table 4 presents an ablation study on CREMA-D and AVE by selectively enabling the SMR and CMD. As a baseline, we use the model trained without both SMR and CMD, which achieves 41.08/15.03 (Multi/Visual) on CREMA-D and 36.11/10.33 on AVE. Overall, SMR mainly improves weak-modality learning by mitigating modality dominance during local training, while its standalone effect on multimodal accuracy can vary across datasets. CMD is most effective when combined with SMR by distilling the aggregated teacher into modality-specific branches, thereby consolidating and amplifying the benefits of local regulation. Quantitatively, on CREMA-D, the full model improves multimodal accuracy by +9.03 points and visual unimodal accuracy by +19.09 points over the baseline, and on AVE, it yields a +0.63 point gain in multimodal accuracy. These results support the complementary roles of the SMR in stabilizing local multimodal optimization and CMD in modality-aware global knowledge transfer. We further examine the dual-head design in SMR (Table 5). The results show that introducing the dual-head consistently improves multimodal accuracy across datasets, indicating that the additional modality-specific supervision helps stabilize multimodal fusion and enhances the effectiveness of dominance-aware regulation.

5 Discussion

Although FedAMB improves performance and stability under heterogeneity, it currently relies on a proxy dataset for distillation, which may be infeasible to obtain in practice and may introduce domain mismatches. Simultaneously, sharing class-wise modality statistics (e.g., prototypes and counts) may reveal coarse information about client-level class distributions under an honest-but-curious server assumption. To mitigate this risk, FedAMB can be combined with standard privacy-enhancing mechanisms, such as secure aggregation, count binning/clipping, and DP-style perturbation of shared prototypes without changing the overall training protocol. Developing proxy-free distillation alternatives and systematically studying privacy utility trade-offs in prototype/logit exchanges remain important directions for future work.

6 Conclusion

To address modality dominance and unimodal fragility in Fed-KD-based MFL, we propose *Federated Adaptive Modality Balancing* (FedAMB). FedAMB is built on two complementary components: *Selective-Modality Regulation* (SMR), which mitigates dominance at its source by selectively regulating biased local dynamics in a state-aware manner, and *Component-wise Modality Distillation* (CMD), which decomposes and regulates global knowledge transfer to prevent the propagation of client-level fusion bias while directly strengthening unimodal representations. Overall, we formulate modality dominance as aggregation-induced contamination of global teacher knowledge and show that the joint application of SMR and CMD stabilizes multimodal fusion under heterogeneous and incomplete modality settings. Extensive experiments on CREMA-D, AVE, and UR-FUNNY demonstrate consistent gains in both multimodal accuracy and unimodal robustness over strong baselines.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (RS-2024-00436934), the ITRC (Information Technology Research Center) grant (RS-2024-00436936), and the Graduate School of Virtual Convergence support program (RS-2023-00254129) supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation).

References

[Adam *et al.*, 2025] Mumin Adam, Abdullatif Albaseer, Uthman Baroudi, and Mohamed Abdallah. Survey of multimodal federated learning: Exploring data integration, challenges, and future directions. *IEEE Open Journal of the Communications Society*, 2025.

[Alfasly *et al.*, 2022] 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.

[Alistarh *et al.*, 2017] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[Cao *et al.*, 2014] 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.

[Che *et al.*, 2023] Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. Multimodal federated learning: A survey. *Sensors*, 23(15):6986, 2023.

[Che *et al.*, 2024] Liwei Che, Jiaqi Wang, Xinyue Liu, and Fenglong Ma. Leveraging foundation models for multimodal federated learning with incomplete modality. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 401–417. Springer, 2024.

[Diao *et al.*, 2021] Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Fan *et al.*, 2023] 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.

[Fan *et al.*, 2024] 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. In *European Conference on Computer Vision*, pages 178–195. Springer, 2024.

[Fraboni *et al.*, 2021] Yann Fraboni, Richard Vidal, and Marco Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International conference on artificial intelligence and statistics*, pages 1846–1854. PMLR, 2021.

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

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

[Kairouz *et al.*, 2021] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[Karimireddy *et al.*, 2020] 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.

[Li and Wang, 2019] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation, 2019.

[Li *et al.*, 2020a] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

[Li *et al.*, 2020b] 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.

[Li *et al.*, 2024] Lin Li, Jianping Gou, Baosheng Yu, Lan Du, and Zhang Yiand Dacheng Tao. Federated distillation: A survey. *arXiv preprint arXiv:2404.08564*, 2024.

[Lin *et al.*, 2018] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[Lin *et al.*, 2020] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.

[Long *et al.*, 2020] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated learning: privacy and incentive*, pages 240–254. Springer, 2020.

[Mammen, 2021] Priyanka Mary Mammen. Federated learning: Opportunities and challenges, 2021.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[Mora *et al.*, 2022] Alessio Mora, Irene Tenison, Paolo Bellavista, and Irina Rish. Knowledge distillation for federated learning: a practical guide. *arXiv preprint arXiv:2211.04742*, 2022.

[Ouyang *et al.*, 2023] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 530–543, 2023.

[Qin *et al.*, 2024] Laiqiao Qin, Tianqing Zhu, Wanlei Zhou, and Philip S Yu. Knowledge distillation in federated learning: A survey on long lasting challenges and new solutions. *arXiv preprint arXiv:2406.10861*, 2024.

[Rieke *et al.*, 2020] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[Sarfraz *et al.*, 2021] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Knowledge distillation beyond model compression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6136–6143. IEEE, 2021.

[Tan *et al.*, 2022] 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.

[Tian *et al.*, 2018] 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.

[Wang *et al.*, 2020] 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.

[Wang *et al.*, 2023] 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.

[Wang *et al.*, 2024] Xingchen Wang, Haoyu Wang, Feijie Wu, Tianci Liu, Qiming Cao, and Lu Su. Towards efficient heterogeneous multi-modal federated learning with hierarchical knowledge disentanglement. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, SenSys '24, page 592–605, New York, NY, USA, 2024. Association for Computing Machinery.

[Wen *et al.*, 2017] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: ternary gradients to reduce communication in distributed deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1508–1518, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Wu *et al.*, 2024] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[Xiong *et al.*, 2023] Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1241–1249, 2023.

[Xu *et al.*, 2021] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of healthcare informatics research*, 5(1):1–19, 2021.

[Yang *et al.*, 2023] Zhiqin Yang, Yonggang Zhang, Yu Zheng, Xinmei Tian, Hao Peng, Tongliang Liu, and Bo Han. Fedfed: Feature distillation against data heterogeneity in federated learning. *Advances in neural information processing systems*, 36:60397–60428, 2023.

[Yin *et al.*, 2021] Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.

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

[Yuan and Ma, 2020] Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020.

[Zadeh *et al.*, 2018] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[Zhang *et al.*, 2021] Tuo Zhang, Chaoyang He, Tianhao Ma, Lei Gao, Mark Ma, and Salman Avestimehr. Federated learning for internet of things. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*, pages 413–419, 2021.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.