# BOOSTING MULTIMODAL FEDERATED LEARNING VIA CHAINED MODALITY OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal Federated Learning (MMFL) enables privacy-preserving collaborative learning across decentralized clients with heterogeneous data and modality availability. However, most existing MMFL methods cast multimodal training as a joint optimization problem, overlooking a key bottleneck: modality competition, where dominant modalities suppress weaker ones and lead to suboptimal global models. To address this, we propose FEDMCHAIN, a balanced MMFL framework that structures federated multimodal training as a chain of modality-wise phases. This phase-wise design gives each modality a dedicated local optimization window on multimodal clients to mitigate modality competition, and further promotes cross-modal complementarity via an error-compensated regularizer. On the server side, we employ a sparse sign-guided aggregation strategy that leverages directional sign agreement for robust intra-modality aggregation, avoids destructive averaging, and supports less frequent synchronization to reduce communication overhead. Extensive experiments on multimodal benchmarks demonstrate that FEDMCHAIN consistently improves predictive performance while requiring less frequent communication than baselines. Code is available at `https://anonymous.4open.science/r/FedMChain-B8D5`.

## 1 INTRODUCTION

Multimodal Federated Learning (MMFL) serves as a privacy-preserving paradigm for collaborative training over distributed multimodal data silos, holding substantial value for domains such as autonomous driving (Zheng et al., 2023) and intelligent healthcare (Orzikulova et al., 2024). Existing MMFL works primarily tackle optimization difficulties arising from *statistical heterogeneity* and *modality heterogeneity* by improving local cross-modal representation alignment (Bao et al., 2023; Yu et al., 2023), designing personalized aggregation and adaptive optimization to accommodate client-specific modality availability and distributions (Chen & Zhang, 2024; Yang et al., 2024; Gao et al., 2025; Pokharel et al., 2025), or jointly optimizing alignment and aggregation for more coherent multimodal collaboration (Qi & Li, 2024; Phung et al., 2025). However, most existing methods implicitly assume that *different modalities follow a relatively balanced optimization process, enabling consistent improvements under joint training*.

The above assumption is often violated in practice by ***modality competition*** (Huang et al., 2022; Peng et al., 2022; Du et al., 2023), a phenomenon empirically observed in centralized multimodal learning. During joint multimodal training, competing optimization signals can bias learning toward dominant modalities, resulting in insufficient optimization of weaker ones. On multimodal clients, such competition can hinder the optimization of weaker modalities, whereas unimodal clients are not subject to the same inter-modality interference. More importantly, the dominant modality may vary across clients, driven by differences in data distribution, data quality, and modality availability. At the *global* level, modality competition can further undermine aggregation stability. Given that the modality competition outcome differs across clients, local models of the same modality may exhibit markedly different convergence progress. Beyond the commonly discussed conflicting update directions in federated learning (FL), such convergence-progress gaps can further complicate effective cross-client knowledge integration. BMSFed (Fan et al., 2024a) is the closest work to our motivation. It mitigates modality competition via modality selection, aggregating only selected client–modality branches in each round. However, this may discard potentially useful information from unselected branches.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To bridge the above gap, we propose FEDMCHAIN, a unified framework for balanced multimodal collaboration in heterogeneous MMFL. Unlike conventional MMFL methods, which jointly optimize all modalities in each communication round, FEDMCHAIN introduces **Modality-Chained Federated Training (MCFT)**, which organizes multimodal training into modality-wise phases and optimizes one modality at a time. This chained schedule provides each modality with an exclusive optimization window under a fixed training-round budget, thereby mitigating client-specific local ***modality competition***. To promote cross-modal complementarity across phases, we further introduce an error-compensated regularizer that emphasizes samples misclassified by preceding modalities. To improve aggregation robustness under heterogeneous client optimization progress, we propose **Sparse Sign-guided Consensus Aggregation (SSCA)**. SSCA leverages sparse directional agreement among client updates to reduce interference caused by magnitude disparities and directional conflicts, while supporting longer synchronization intervals and thus lower communication frequency. In summary, our contributions are threefold:

- We propose FEDMCHAIN, a novel MMFL framework that performs sequential modality optimization in heterogeneous federated settings, facilitating sufficient modality-specific optimization and alleviating modality competition.
- We introduce a sparse sign-guided aggregation strategy that leverages directional sign agreement for robust intra-modality integration, with reduced communication overhead.
- Extensive experiments on three widely used multimodal benchmarks demonstrate the superior effectiveness and efficiency of the proposed method.

## 2 RELATED WORK

### 2.1 MULTIMODAL FEDERATED LEARNING

Based on whether clients share a consistent modality configuration, MMFL can be categorized as consistent and inconsistent MMFL (Che et al., 2023). We focus on the inconsistent setting, where modality availability varies across clients. Existing approaches for inconsistent MMFL mainly fall into two lines: 1) *Modularity-based training and aggregation*. These approaches treat a part of the client models as shared modules for federated aggregation, thereby facilitating knowledge sharing among clients (Yang et al., 2022; Zhang et al., 2023b; Yuan et al., 2024; Cho et al., 2022; Qi & Li, 2024; Li et al., 2024; Chen & Zhang, 2024; Phung et al., 2025; Pokharel et al., 2025; Ouyang et al., 2023). 2) *Representation-based aggregation*. Such approaches utilize prototypes, intermediate representations, or adaptive representation controls as carriers of knowledge, aggregating them on the server to guide the local training process (Yu et al., 2023; Zeng et al., 2024; Le et al., 2024; Guo et al., 2024; Gao et al., 2025; Seo et al., 2025; Nguyen et al., 2025). However, most existing inconsistent MMFL methods retain joint multimodal optimization on multimodal clients and are not explicitly designed to address modality competition during local training. Notably, BMSFed (Fan et al., 2024a) targets modal bias by selecting modality-specific networks for communication and aggregating global prototypes to strengthen weaker modalities; yet its selection mechanism may underutilize complementary information and does not explicitly decouple inter-modality optimization during local joint training.

### 2.2 IMBALANCED MULTIMODAL LEARNING

Multimodal learning often suffers from modality competition (Wang et al., 2020; Peng et al., 2022), where optimization can become biased toward dominant modalities, leading to insufficient learning of weaker ones. Recent approaches address this issue from multiple angles. Some methods (Wang et al., 2020; Peng et al., 2022; Fan et al., 2023; Li et al., 2023; Wei & Hu, 2024) adjust the gradient magnitudes to slow down the learning of dominant modalities, thereby maintaining a more balanced optimization process across modalities. Others (Wu et al., 2022; Du et al., 2023; Zhang et al., 2023a) introduce additional auxiliary modules to explicitly reweight or recalibrate the contribution of each modality. Another line of work explores improving training paradigms (Fan et al., 2024b; Hua et al., 2024; Jiang et al., 2024; Zhang et al., 2024) or leveraging data augmentation (Hwang et al., 2025; Ma et al., 2025) to achieve more coordinated multimodal representation learning. However, directly applying these methods to MMFL is challenging: gradient modulation and auxiliary recalibration add computational overhead and rely on sufficiently rich local data, while alternating training can

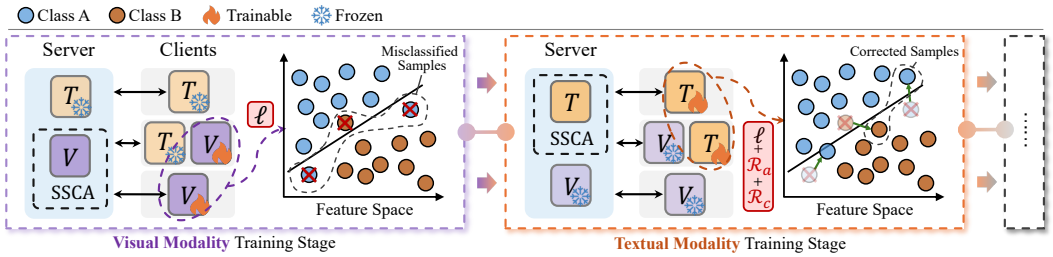



Figure 1: Illustration of Modality-Chained Federated Training (MCFT), using visual ($V$) and textual ($T$) modalities as an example.

markedly prolong local training, hurting system efficiency in communication-limited federated settings.

## 3 METHOD

In this section, we propose the FEDMCHAIN framework, comprising Modality-Chained Federated Training (MCFT) and Sparse Sign-guided Consensus Aggregation (SSCA), as shown in Figure 1 and Figure 2, respectively. Further details of FEDMCHAIN are provided below.

### 3.1 PROBLEM FORMULATION

We consider a MMFL framework consisting of a central server and a set of clients $\mathcal{C} = \{1, 2, \ldots, |\mathcal{C}|\}$. Each client $i \in \mathcal{C}$ holds a local dataset $\mathcal{D}_i = \left\{(\{x_{i,j}^{(m)}\}_{m\in\mathcal{M}_i}, y_{i,j})\right\}_{j=1}^{n_i}$, where $n_i = |\mathcal{D}_i|$ is the number of local samples and $N = \sum_{i=1}^{|\mathcal{C}|} n_i$ is the total number of samples. Here, $\mathcal{M}_i \subseteq \mathcal{M}$ denotes the subset of modalities available to client $i$, and $\mathcal{M}$ is the global modality set. Each $x_{i,j}^{(m)}$ is the input of modality $m$ for sample $j$ on client $i$, and $y_{i,j}$ is the corresponding label. Clients differ in their modality configurations: some are unimodal ($|\mathcal{M}_i| = 1$), while others are multimodal ($|\mathcal{M}_i| > 1$). Each client $i$ maintains modality-specific encoders and classifiers with parameters $\Theta_i = \{(\theta_{E,i}^{(m)}, \theta_{C,i}^{(m)}) \mid m \in \mathcal{M}_i\}$. Given input $x_{i,j}^{(m)}$, the modality prediction score is $z_{i,j}^{(m)} = f_{C,i}^{(m)}(f_{E,i}^{(m)}(x_{i,j}^{(m)}))$. Multimodal clients may fuse predictions from multiple modalities using a fusion operator $\mathcal{F}_i(\cdot)$, while unimodal clients rely on a single branch. The server optimizes modality-specific global branches $\Theta_G = \{(\theta_{E,G}^{(m)}, \theta_{C,G}^{(m)}) | m \in \mathcal{M}\}$. Given these branches, the empirical risk of client $i$ is

$$\mathcal{L}_i(\Theta_i; D_i) \triangleq \frac{1}{n_i} \sum_{j=1}^{n_i} \ell\Big(\mathcal{F}_i\Big(\{z_{i,j}^{(m)}\}_{m\in\mathcal{M}_i}\Big), y_{i,j}\Big), \tag{1}$$

and the global objective is

$$\min_{\Theta_G} F(\Theta_G), \quad F(\Theta_G) := \sum_{i=1}^{|\mathcal{C}|} w_i \mathcal{L}_i(\Theta_G^{\mathcal{M}_i}; D_i), \tag{2}$$

where $w_i = \frac{n_i}{N}$, $\Theta_G^{\mathcal{M}_i}$ denotes the subset of global modality branches available to client $i$ and $\ell(\cdot, \cdot)$ denotes the per-sample local loss function.

### 3.2 MODALITY-CHAINED FEDERATED TRAINING (MCFT)

Prior work (Wang et al., 2020; Peng et al., 2022; Fan et al., 2023; Du et al., 2023) attributes modality competition to gradient conflict across modalities under joint training, i.e., simultaneously optimizing multiple modality-specific branches under a shared objective. Let $z_{i,j}^{\mathrm{f}} = \mathcal{F}_i(\{z_{i,j}^{(m)}\}_{m\in\mathcal{M}_i})$ denote the fused prediction scores for the $j$-th sample at client $i$. Following the analysis proposed by Hua et al. (Hua et al., 2024), the learning behavior can be characterized through a gradient alignment

analysis: a modality receives a favorable optimization signal when the gradient of its modality-specific prediction is directionally consistent with that of the fused prediction,

$$\left\langle \nabla_{z_{i,j}^{(m)}} \ell,\ \nabla_{z_{i,j}^{\mathrm{f}}} \ell \right\rangle > 0, \tag{3}$$

In contrast, persistent gradient misalignment ($< 0$) can hinder the optimization progress of the corresponding modality, contributing to imbalanced learning across modalities.

In MMFL, modality competition not only affects client-side optimization, but can also induce heterogeneous optimization progress across clients, complicating the aggregation of same-modality updates at the server. To mitigate client-side modality competition, we propose Modality-Chained Federated Training (MCFT), a *modality-by-modality* optimization paradigm that updates one modality branch at a time, thereby reducing inter-modality gradient interference. Concretely, MCFT decomposes global training into a sequence of modality-specific phases. In phase $m$, only parameters associated with modality $m$ are optimized:

a) **Unimodal clients** with modality $m$ update their modality-$m$ branch locally;

b) **Multimodal clients** update only modality-$m$ parameters while freezing all other modality branches;

c) The **server** periodically aggregates updates restricted to modality $m$ from participating clients to maintain modality-consistent global optimization.

Nevertheless, phase-wise modality-specific optimization alone is insufficient to fully leverage multimodal collaboration. While decoupling modality optimization mitigates gradient interference, effective MMFL requires (i) *cross-modal alignment* to induce consistent semantics across modalities and (ii) *cross-modal complementarity* to ensure that distinct modalities contribute non-redundant information. To this end, we augment the client objective in each phase with explicit regularizers targeting both properties. Formally, during the phase with active modality $m$, client $i$ minimizes the following mini-batch objective:

$$\begin{aligned}\mathcal{L}_i^{(m)} =& \mathbb{E}_{\mathcal{B}_i \sim \mathcal{D}_i} \Bigg[ \frac{1}{|\mathcal{B}_i|} \sum_{(x_{i,b}, y_{i,b}) \in \mathcal{B}_i} \underbrace{\ell\Big(z_{i,b}^{(m)},\, y_{i,b}\Big)}_{\text{Task}} \\ &+ \mathbb{I}[\,|\mathcal{P}_m \cap \mathcal{M}_i| > 0\,] \Big( \underbrace{\lambda_a \mathcal{R}_{a,i}^{(m)}(\mathcal{B}_i)}_{\text{Alignment}} + \underbrace{\frac{\lambda_c}{|\mathcal{B}_i|} \sum_{(x_{i,b}, y_{i,b}) \in \mathcal{B}_i} \mathcal{R}_{c,i,b}^{(m)}}_{\text{Complementarity}} \Big) \Bigg],\end{aligned} \tag{4}$$

where $\mathcal{P}_m$ denotes the set of modalities preceding $m$ in the chain, $\mathcal{R}_a$ and $\mathcal{R}_c$ are regularizers for cross-modal alignment and complementarity, respectively; $\mathbb{I}[\cdot]$ enforces that these terms are applied only when the required preceding modalities are available to client $i$.

**(i) Cross-modal Alignment Regularizer.** We encourage the active modality $m$ to learn modality-invariant semantics by aligning its representation with those of modalities in $\mathcal{P}_m \cap \mathcal{M}_i$. Let $h_{i,b}^{(m)} = f_{E,i}^{(m)}(x_{i,b}^{(m)})$ denote the latent feature of modality $m$ for the $b$-th sample in a mini-batch $\mathcal{B}_i$ of size $B = |\mathcal{B}_i|$. We $\ell_2$-normalize features as $\tilde{h} = h/\|h\|_2$ and define an alignment regularizer that pulls together cross-modal representations of the same sample while pushing apart mismatched pairs within the mini-batch:

$$\begin{aligned}\mathcal{R}_{a,i}^{(m)}(\mathcal{B}_i) =& \frac{1}{|\mathcal{P}_m \cap \mathcal{M}_i|} \sum_{m' \in \mathcal{P}_m \cap \mathcal{M}_i} \\ & \left[ -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\exp\big(\mathcal{S}(\tilde{h}_{i,b}^{(m)}, \tilde{h}_{i,b}^{(m')})/\tau\big)}{\sum_{b'=1}^{B} \exp\big(\mathcal{S}(\tilde{h}_{i,b}^{(m)}, \tilde{h}_{i,b'}^{(m')})/\tau\big)} \right],\end{aligned} \tag{5}$$

where $\mathcal{S}(\cdot,\cdot)$ denotes cosine similarity, $\tau$ is a temperature coefficient, and $b, b'$ index samples within the mini-batch.

**(ii) Cross-modal Complementarity Regularizer.** To promote cross-modal knowledge complementarity, we introduce an error-compensation regularizer that increases the loss contribution of samples

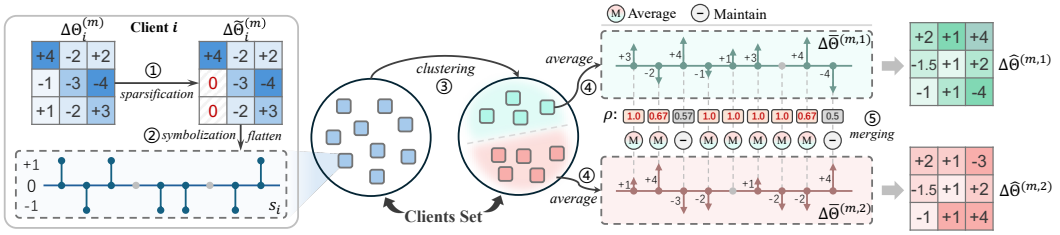



Figure 2: Illustration of Sparse Sign-guided Consensus Aggregation (SSCA). We use $K = 2$ as an example.

that are predicted with low confidence by modalities trained in preceding phases. Specifically, predictions from preceding modalities are used to identify samples that remain difficult or misclassified, assigning them larger weights when optimizing the current modality. This encourages the current modality to capture complementary information that has not been sufficiently exploited by preceding modalities. For client $i$ during the modality-$m$ phase, we first aggregate the logits from preceding modalities available locally:

$$\bar{p}_{i,j}^{(<m)}(y_{i,j} \mid x_{i,j}) = \mathrm{softmax}\Big(\sum_{m'} z_{i,j}^{(m')}\Big)_{y_{i,j}}, \tag{6}$$

where $m' \in \mathcal{P}_m \cap \mathcal{M}_i$ and $\bar{p}_{i,j}^{(<m)}(y_{i,j} \mid x_{i,j})$ denotes the aggregated confidence on the ground-truth label. We then define an error-compensation weight from the preceding-phase aggregated confidence:

$$e_{i,j}^{(m)} \;=\; 1 - \bar{p}_{i,j}^{(<m)}(y_{i,j} \mid x_{i,j})\,, \tag{7}$$

and construct the complementarity regularizer by modulating the supervised loss of the active modality $m$:

$$\mathcal{R}_{c,i,j}^{(m)} \;=\; e_{i,j}^{(m)}\, \ell\Big(z_{i,j}^{(m)},\, y_{i,j}\Big)\,. \tag{8}$$

### 3.3 Sparse Sign-guided Consensus Aggregation (SSCA)

Although MCFT mitigates modality competition through modality-wise optimization, heterogeneous clients may still exhibit different levels of optimization progress within the same modality, leading to inconsistent local updates. Directly averaging these updates may result in destructive interference and compromise global optimization. To address this issue, we propose Sparse Sign-guided Consensus Aggregation (SSCA), which leverages directional agreement to selectively consolidate intra-modality updates, reducing destructive averaging under cross-client update heterogeneity.

For an active modality $m$, SSCA maintains $K$ cluster-specific server models $\{\Theta_G^{(m,k)}\}_{k=1}^{K}$. Each participating client $i$ uses its current cluster index $c_i^{(m)}$ to compute $\Delta\Theta_i^{(m)} = \Theta_i^{(m)} - \Theta_G^{(m,c_i^{(m)})}$. We retain the largest-magnitude coordinates with retention ratio $\kappa$ and extract their signs:

$$s_i = \mathrm{sign}\big(\mathcal{T}(\Delta\Theta_i^{(m)}, \kappa)\big). \tag{9}$$

The flattened sign vectors are clustered by KMeans into $K$ directional groups $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, and the resulting labels update $c_i^{(m)}$ within the current modality stage. Clustering is performed independently across modality stages, so cluster identities are not shared across modalities.

Within each cluster $\mathcal{C}_k$, we aggregate the member updates to form a cluster-level consensus update that captures their common optimization direction. Given the sparsified modality update $\Delta\tilde{\Theta}_i^{(m)} = \mathcal{T}(\Delta\Theta_i^{(m)}, \kappa)$, we compute the cluster consensus by a sample-weighted mean:

$$\Delta\bar{\Theta}^{(m,k)} = \frac{\sum_{i\in\mathcal{C}_k} w_i\, \Delta\tilde{\Theta}_i^{(m)}}{\sum_{i\in\mathcal{C}_k} w_i}. \tag{10}$$

We then coordinate the $K$ cluster consensuses through a parameter consistency score $\rho$, which measures cross-cluster directional agreement at each parameter coordinate and determines whether to

unify the corresponding updates or preserve specific differences. Concretely, for each coordinate $q$, let $\Delta\bar{\Theta}_q^{(m,k)}$ denote the $q$-th entry of $\Delta\bar{\Theta}^{(m,k)}$. We define $\rho_q$ as the dominance ratio of the stronger direction:

$$p_q = \sum_{k:\Delta\bar{\Theta}_q^{(m,k)}>0} \alpha_k \big|\Delta\bar{\Theta}_q^{(m,k)}\big|, \;\; n_q = \sum_{k:\Delta\bar{\Theta}_q^{(m,k)}<0} \alpha_k \big|\Delta\bar{\Theta}_q^{(m,k)}\big|, \tag{11}$$

$$\rho_q = \frac{\max(p_q, n_q)}{p_q + n_q + \varepsilon}. \tag{12}$$

Here, $\varepsilon$ is a small constant for numerical stability, $\alpha^k = \sum_{i\in\mathcal{C}_k} w_i$ is the total weight of cluster $\mathcal{C}_k$. The score $\rho_q \in [0,1]$ quantifies how strongly one update direction dominates across clusters at coordinate $q$: values close to 1 indicate that most cluster-level consensuses agree on the same sign, while smaller values imply severe sign conflicts and thus low cross-cluster consistency.

Based on $\rho_q$, SSCA applies a thresholding rule with hyper-parameter $\pi \in (0,1)$ to decide whether to merge cluster updates at coordinate $q$. Specifically, we construct a coordinate-wise cross-cluster mask

$$\mu_q = \mathbb{I}(\rho_q \geq \pi)\,. \tag{13}$$

For $\mu_q = 1$, we first identify the dominant direction

$$d_q = \mathrm{sign}(p_q - n_q), \tag{14}$$

and aggregate only the cluster entries aligned with $d_q$:

$$\Delta\hat{\Theta}_q^{(m)} = \frac{\sum_{k=1}^K \alpha^k\, \Delta\bar{\Theta}_q^{(m,k)}\, \mathbb{I}\Big(\mathrm{sign}(\Delta\bar{\Theta}_q^{(m,k)}) = d_q\Big)}{\sum_{k=1}^K \alpha^k\, \mathbb{I}\Big(\mathrm{sign}(\Delta\bar{\Theta}_q^{(m,k)}) = d_q\Big) + \varepsilon}, \tag{15}$$

For $\mu_q = 0$ (i.e., $\rho_q < \pi$), we do not enforce cross-cluster merging and simply retain $\Delta\hat{\Theta}_q^{(m,k)} = \Delta\bar{\Theta}_q^{(m,k)}$. Combining the two cases, the final aggregated update can be expressed compactly as

$$\Delta\hat{\Theta}_q^{(m,k)} = \mu_q\, \Delta\hat{\Theta}_q^{(m)} + (1-\mu_q)\, \Delta\bar{\Theta}_q^{(m,k)}. \tag{16}$$

This rule unifies clusters only on coordinates with clear directional dominance, while keeping sign-ambiguous coordinates at the cluster-level consensuses to avoid interference from conflicting optimization directions.

Finally, the server updates the modality-specific global parameters via a controlled merge step:

$$\Theta_{G,r+1}^{(m,k)} = \Theta_{G,r}^{(m,k)} + \lambda_{\mathrm{merge}} \Delta\hat{\Theta}^{(m,k)}, \tag{17}$$

where $\lambda_{\mathrm{merge}} \in (0,1]$ regulates the integration strength of the aggregated update to stabilize optimization under heterogeneous client updates. Thus, SSCA selectively shares information across clusters only where sufficient directional consensus exists, while preserving cluster-specific updates on conflicting coordinates.

Notably, SSCA reconciles divergent client updates by emphasizing directional consensus during aggregation, which improves robustness when client updates are misaligned under heterogeneity. Empirically, this robustness makes SSCA less sensitive to longer synchronization intervals (Table 5), allowing the server to aggregate less frequently and thus reducing communication overhead.

### 3.4 Convergence Analysis

We provide a stage-wise convergence analysis of FedMChain in smooth non-convex federated optimization. For a fixed modality stage $m$, we condition on a fixed cluster partition and consider the corresponding objective $F^{(m,k)}$ induced by clients in $\mathcal{C}_k$. Under the assumptions in Appendix A.3, including descent correlation and a bounded conditional second moment of the SSCA update, Theorem A.5 establishes the non-convex stationarity bound

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\big[\|\nabla F^{(m,k)}(\Theta_r)\|^2\big] \leq \mathcal{O}\left(\frac{1}{R}\right) + \mathcal{O}(\sigma_{\mathrm{agg}}^2), \tag{18}$$

where $R$ is the number of server aggregation rounds and $\sigma_{\mathrm{agg}}^2$ captures the SSCA-induced aggregation noise. ***The details of the proof are provided in Appendix A.3.***

Table 1: Performance comparison on CREMA-D, AVE, and CMU-MOSEI. We report overall accuracy (ACC) averaged over all clients; $\text{ACC}_v$, $\text{ACC}_a$, and $\text{ACC}_t$ averaged over all clients containing the visual, audio, and text modality, respectively, including both unimodal and multimodal clients; and $\text{ACC}_m$ averaged over multimodal clients. Best results are in **bold** and the second-best results are underlined.

| Method | CREMA-D | | | | AVE | | | | CMU-MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | $\text{ACC}_v$ | $\text{ACC}_a$ | $\text{ACC}_m$ | ACC | $\text{ACC}_v$ | $\text{ACC}_a$ | $\text{ACC}_m$ | ACC | $\text{ACC}_v$ | $\text{ACC}_a$ | $\text{ACC}_t$ | $\text{ACC}_m$ |
| Local | 48.96 | 29.64 | 50.45 | 50.69 | 45.14 | 26.37 | 43.81 | 49.25 | 55.48 | 40.36 | 39.15 | 59.84 | 54.15 |
| FedAvg | 47.11 | 31.33 | 49.81 | 51.75 | 47.18 | 29.49 | 44.06 | 50.24 | 55.63 | 41.22 | 39.84 | 58.15 | 55.39 |
| FedProx | 48.35 | 30.74 | 50.69 | 51.97 | 48.25 | 30.26 | 44.68 | 51.42 | 56.59 | 40.03 | 40.24 | 60.11 | 55.45 |
| SCAFFOLD | 48.27 | 31.41 | 51.25 | 50.85 | 51.09 | 30.57 | 48.23 | 53.75 | 58.83 | 41.62 | 41.47 | 62.06 | 57.31 |
| FedMSplit | 51.62 | 33.59 | 51.02 | 53.46 | 55.16 | 29.64 | 50.18 | 58.83 | 61.17 | 43.74 | 44.12 | 64.52 | 60.92 |
| CreamFL | 53.66 | 35.68 | 52.95 | 55.13 | 55.44 | 32.25 | 53.46 | 58.98 | - | - | - | - | - |
| Harmony | 55.49 | 40.54 | 51.89 | 57.92 | 56.85 | 34.41 | 55.06 | 60.86 | 62.59 | 44.18 | 46.36 | 66.21 | 63.45 |
| HAMFL | 52.53 | 41.16 | 52.21 | 54.60 | 56.17 | 33.51 | 53.78 | 59.45 | 62.40 | 43.89 | 46.08 | 65.97 | 63.83 |
| $\text{M}^3$Fed | 53.94 | 40.15 | 51.52 | 55.98 | 55.28 | 33.05 | 54.12 | 59.36 | 61.53 | 45.14 | 48.26 | **68.46** | 63.77 |
| FedMBridge | 54.88 | 39.82 | 51.95 | 56.75 | 57.33 | 35.69 | 56.33 | 61.84 | 61.29 | 45.96 | **50.38** | 68.21 | 64.68 |
| PEPSY | 56.85 | 42.37 | 53.46 | 59.32 | 57.25 | 36.98 | 58.03 | 61.29 | 63.64 | 45.77 | 49.52 | 68.35 | 65.22 |
| BMSFed | 56.14 | **44.25** | 53.65 | 58.69 | 57.62 | 37.96 | **58.25** | 60.22 | 63.05 | **46.24** | 49.76 | 68.12 | 64.30 |
| Ours | **58.36** | 43.89 | **54.10** | **60.68** | **59.85** | **38.49** | 57.95 | **62.74** | **64.96** | 45.86 | 50.25 | 68.37 | **65.75** |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets.** We evaluate our proposed method on three commonly used multimodal benchmarks: **CREMA-D** (Cao et al., 2014), **AVE** (Tian et al., 2018), and **CMU-MOSEI** (Zadeh et al., 2018). To simulate federated multimodal settings with heterogeneous data distributions, we partition each dataset into multiple clients using a Dirichlet distribution with concentration parameter $\beta \in \{0.5, 1.0\}$[1],where a smaller $\beta$ indicates higher heterogeneity. Within each client, the local data are split into 80% for training and 20% for testing. Additionally, we simulate heterogeneous modality availability by masking modalities for a subset of clients to create diverse cross-client modality configurations. ***Further dataset and partitioning details are included in Appendix A.1.1***.

**Implementation Details.** All methods are implemented in PyTorch (Paszke et al., 2017) and trained on 8 NVIDIA RTX 3090 GPUs. For a fair comparison, all baselines are evaluated under the same data partition protocol and identical backbone/model configurations. The local batch size is set to 128, and each training round runs one local epoch. All methods are trained for 500 training rounds. The baselines perform server aggregation after every round, whereas FEDMCHAIN aggregates every 20 training rounds. For our method, we set the retention ratio $\kappa$ to 0.7 for CREMA-D and AVE and 0.8 for CMU-MOSEI, the consensus threshold $\pi$ to 0.8, the merge learning rate $\lambda_{\text{merge}}$ to 0.9, and the loss weighting coefficients $\lambda_a$ and $\lambda_c$ to 0.4 and 1.0, respectively. For all datasets, we repeat each experiment three times with different random seeds, and report the mean performance. ***Further details of the experimental setup are given in Appendix A.1.2.***

**Baselines.** We compare our method with **Local** training, classical FL methods (**FedAvg** (McMahan et al., 2017), **FedProx** (Li et al., 2020), and **SCAFFOLD** (Karimireddy et al., 2020)), and representative heterogeneous MMFL approaches, including **Harmony** (Ouyang et al., 2023), **FedMSplit** (Chen & Zhang, 2022), **CreamFL** (Yu et al., 2023), **HAMFL** (Qi & Li, 2024), **$\text{M}^3$Fed** (Li et al., 2024), **FedMBridge** (Chen & Zhang, 2024), **PEPSY** (Nguyen et al., 2025), and **BMSFed** (Fan et al., 2024a). These baselines cover modality heterogeneity, missing modalities, and modality competition, with BMSFed being particularly relevant as it explicitly targets modality competition in MMFL.

[1]Unless otherwise specified, all results reported in this paper are obtained with $\beta = 1.0$.

### 4.2 PERFORMANCE ANALYSIS

Table 1 summarizes the performance on the three datasets. Our method consistently achieves the best overall ACC and multimodal-client $\text{ACC}_m$. Compared with the strongest baseline for each metric, it improves ACC by 1.51, 2.23, and 1.32 points, and $\text{ACC}_m$ by 1.36, 0.90, and 0.53 points on CREMA-D, AVE, and CMU-MOSEI, respectively, demonstrating its effectiveness under heterogeneous modality distributions. Notably, these overall gains are accompanied by competitive performance across different modality groups. Although our method does not achieve the best result on every modality-wise metric, it achieves the best $\text{ACC}_a$ on CREMA-D and $\text{ACC}_v$ on AVE, while remaining close to the best results on the other modality-wise metrics. This indicates that FEDMCHAIN maintains strong performance across clients with different modality availability while effectively exploiting multimodal collaboration. Recent heterogeneous MMFL methods, such as Harmony, FedMBridge, PEPSY, and BMSFed, generally outperform classical FL baselines, highlighting the importance of explicitly handling modality heterogeneity. Nevertheless, our method maintains a clear advantage in both overall and multimodal-client performance. In particular, BMSFed achieves strong performance on several modality-wise metrics by explicitly addressing modality competition, whereas our higher ACC and $\text{ACC}_m$ indicate a better overall balance between performance across heterogeneous clients and multimodal collaboration. In contrast, classical FL methods can even underperform Local training (e.g., FedAvg on CREMA-D), showing that naive aggregation under heterogeneous multimodal data may induce negative transfer.

### 4.3 ABLATION STUDIES

Table 2 isolates the effect of each component. Removing MCFT causes the largest performance degradation, showing that modality-wise training is a major contributor to the overall performance. Disabling SSCA also consistently reduces accuracy. Without SSCA, server aggregation reduces to plain averaging applied separately to each modality branch, indicating that directional-consensus aggregation provides clear benefits over plain averaging under heterogeneous client updates. Removing either $R_a$ or $R_c$ also degrades performance, while removing both causes a larger drop than removing either term alone. This is consistent with their complementary roles: $R_a$ promotes cross-modal alignment, whereas $R_c$ encourages subsequent modalities to focus on samples that are difficult for preceding modalities.

Table 2: Ablation study of our framework on CREMA-D, AVE, and CMU-MOSEI.

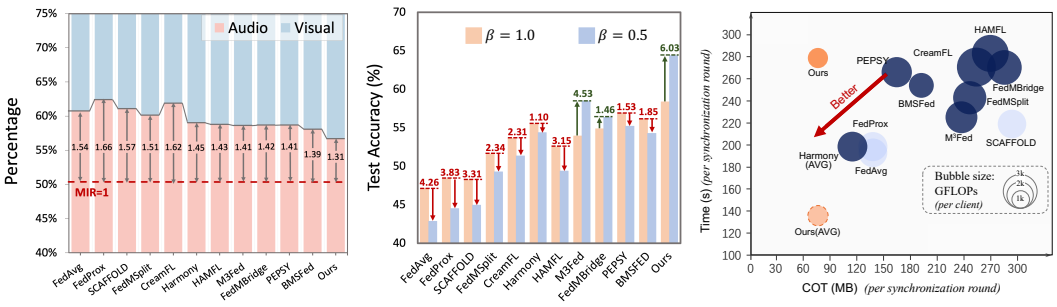
| Method | CREMA-D | AVE | CMU-MOSEI |
|---|---|---|---|
| w/o MCFT | 54.96 ↓3.40 | 54.82 ↓5.03 | 60.39 ↓4.57 |
| w/o SSCA | 56.25 ↓2.11 | 55.43 ↓4.42 | 60.98 ↓3.98 |
| w/o $\mathcal{R}_a$ | 57.12 ↓1.24 | 58.14 ↓1.71 | 64.25 ↓0.71 |
| w/o $\mathcal{R}_c$ | 57.03 ↓1.33 | 57.29 ↓2.56 | 63.84 ↓1.12 |
| w/o $\mathcal{R}_a$&$\mathcal{R}_c$ | 56.31 ↓2.05 | 55.60 ↓4.25 | 63.08 ↓1.88 |
| Ours | **58.36** | **59.85** | **64.96** |

### 4.4 FURTHER ANALYSIS

**Modality Competition.** Modality competition often manifests as a performance imbalance across modalities. We therefore report the modality imbalance ratio (MIR) as a performance-based measure of this imbalance, where a larger value indicates greater disparity across modalities (see Appendix A.2 for the formal definition). As shown in Figure 3a on CREMA-D, FEDMCHAIN achieves the lowest MIR among all compared approaches (1.31). Together with its strong overall and modality-wise performance in Table 1, this result indicates that FEDMCHAIN achieves a more balanced utilization of different modalities while maintaining competitive predictive performance.

**Impact of Modality Training Orders.** Table 3 evaluates FEDMCHAIN under different modality orders. Although ACC varies across orders, FEDMCHAIN remains competitive across the tested schedules, with several orders outperforming BMSFed. This order sensitivity is consistent with the sequential design of MCFT, where preceding modalities provide reference signals for subsequent stages and can therefore influence later optimization. More importantly, FEDMCHAIN consistently achieves lower MIR than BMSFed under every order. Thus, while ordering affects the resulting predictive performance, the consistently lower MIR indicates reduced modality imbalance across different training orders, supporting the effectiveness of FEDMCHAIN in mitigating modality competition.

**Degree of Heterogeneity of Data Distribution.** We further evaluate all methods on CREMA-D under two levels of statistical heterogeneity ($\beta \in \{0.5, 1.0\}$), where a smaller $\beta$ corresponds to a




(a) MIR (b) (c)

Figure 3: (a) Modality imbalance ratio (MIR) on CREMA-D. (b) Effect of different levels of data heterogeneity on the performance of the compared methods. (c) Computation–communication trade-offs on CREMA-D (x-axis: communication cost per synchronization round; y-axis: average runtime per synchronization round; bubble size: local computation per client).

Table 3: Effect of different modality training orders on ACC and MIR.

| Metric | CREMA-D | | | AVE | | | CMU-MOSEI | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a→v | v→a | BMSFed | a→v | v→a | BMSFed | t→a→v | t→v→a | a→v→t | a→t→v | v→a→t | v→t→a | BMSFed |
| ACC ↑ | 58.36 | 56.09 | 56.14 | 59.85 | 56.33 | 57.62 | 64.96 | 64.58 | 62.16 | 63.24 | 62.07 | 62.37 | 63.05 |
| MIR ↓ | 1.31 | 1.34 | 1.39 | 1.58 | 1.62 | 1.74 | 1.82 | 1.85 | 1.93 | 1.86 | 1.95 | 1.88 | 2.25 |

more non-IID data distribution. As shown in Figure 3b, classical FL baselines generally deteriorate under the stronger heterogeneous setting, whereas heterogeneous MMFL methods exhibit greater robustness with method-dependent variations. FEDMCHAIN achieves the highest accuracy under both settings and maintains strong performance at $\beta = 0.5$, demonstrating robustness across different degrees of statistical heterogeneity. This result is consistent with SSCA, which exploits directional agreement among client updates during aggregation.

**Efficiency Analysis.** Figure 3c compares the computation–communication trade-offs of the federated methods. FEDMCHAIN requires the lowest per-client local computation and substantially reduces the transmitted payload because MCFT updates and communicates only the active modality branch while freezing the others. SSCA introduces additional server-side aggregation cost, which is reflected by the runtime of synchronization rounds. However, its robustness to longer synchronization intervals allows this cost to be amortized over multiple local-training rounds. We therefore additionally report Ours(AVG), which averages the runtime over both synchronization and non-synchronization rounds. Under the same total training-round budget, Ours(AVG) achieves the lowest average runtime among the compared methods, demonstrating a favorable overall computation–communication trade-off.

***Due to space limitations, additional experimental results are provided in Appendix A.2.***

## 5 Conclusion & Limitation

In this work, we identify modality competition as an important optimization challenge in heterogeneous MMFL and propose FEDMCHAIN. By decomposing joint multimodal optimization into modality-wise training stages, FEDMCHAIN enables sufficient optimization of individual modalities while preserving cross-modal collaboration through alignment and error-compensation regularization. SSCA further reduces interference among heterogeneous client updates by exploiting their directional agreement. Extensive experiments on three multimodal benchmarks demonstrate improvements in overall and multimodal-client performance, reduced modality imbalance, and a favorable computation–communication trade-off. A limitation of the current framework is that the modality training order is fixed in advance rather than adapted to the evolving optimization state. Looking ahead, an interesting direction is to develop adaptive modality-ordering strategies that dynamically determine the training sequence according to modality characteristics and optimization dynamics, further enhancing the flexibility of FEDMCHAIN across heterogeneous federated scenarios.

### AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing, including grammar correction and improvements to clarity and readability. They were not used to generate research ideas, design the proposed method, conduct experiments, derive theoretical results, or produce scientific claims. All AI-assisted edits were carefully reviewed by the authors, who take full responsibility for the final content of this work.

### ETHICS STATEMENT

This work does not involve new data collection or direct interaction with human subjects. All experiments are conducted on existing multimodal research benchmarks, including CREMA-D, AVE, and CMU-MOSEI. The federated learning setting is simulated through client-wise data partitioning and modality masking, without collecting private user data. As with other learning-based methods, the resulting models may inherit biases present in the original datasets. This work is intended for research purposes.

### REPRODUCIBILITY STATEMENT

We provide detailed dataset descriptions, preprocessing procedures, model architectures, and training configurations in Appendix A.1. Additional experimental results and hyperparameter analyses are reported in Appendix A.2, while the assumptions and convergence analysis are provided in Appendix A.3. We also provide anonymous source code to facilitate reproduction of our results.

## REFERENCES

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv preprint arXiv:2312.13508*, 2023.

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.

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

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*, pp. 87–96, 2022.

Jiayi Chen and Aidong Zhang. Fedmbridge: Bridgeable multimodal federated learning. In *Forty-first International Conference on Machine Learning*, 2024.

Yae Jee Cho, Andre Manoel, Gauri Joshi, Robert Sim, and Dimitrios Dimitriadis. Heterogeneous ensemble knowledge transfer for training large models in federated learning. *arXiv preprint arXiv:2204.12703*, 2022.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, 2014. doi: 10.1109/ICASSP.2014.6853739.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

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*, pp. 8632–8656. PMLR, 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*, pp. 20029–20038, 2023.

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, 2024a. URL https://arxiv.org/abs/2401.00403.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5470–5478, 2024b.

Min Gao, Haifeng Zheng, Xinxin Feng, and Ran Tao. Multimodal fusion using multi-view domains for data heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16736–16744, 2025.

Qi Guo, Minghao Yao, Zhen Tian, Saiyu Qi, Yong Qi, Yun Lin, and Jin Song Dong. Contribution evaluation of heterogeneous participants in federated learning via prototypical representations. *arXiv preprint arXiv:2407.02073*, 2024.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*, 2024.

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), 2022. URL https://arxiv.org/abs/2203.12221.

Seong-Hyeon Hwang, Soyoung Choi, and Steven Euijong Whang. Midas: Misalignment-based data augmentation strategy for imbalanced multimodal learning, 2025. URL https://arxiv.org/abs/2509.25831.

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang. Multimodal classification via modal-aware interactive enhancement. *arXiv preprint arXiv:2407.04587*, 2024.

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*, pp. 5132–5143. PMLR, 2020.

Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *arXiv preprint arXiv:2401.13898*, 2024.

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*, pp. 22214–22224, 2023.

Shuai Li, Fan Qi, Zixin Zhang, and Changsheng Xu. Cross-modal meta consensus for heterogeneous federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 975–984, 2024.

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.

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hDH3KfZSsF.

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*, pp. 1273–1282. PMLR, 2017.

Duong M. Nguyen, Trong Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Adiba Orzikulova, Jaehyun Kwak, Jaemin Shin, and Sung-Ju Lee. Federated learning for time-series healthcare sensing with incomplete modalities. *arXiv preprint arXiv:2405.11828*, 2024.

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*, pp. 530–543, 2023.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

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*, pp. 8238–8247, 2022.

Thu Hang Phung, Duong M. Nguyen, Thanh Trung Huynh, Quoc Viet Hung Nguyen, Trong Nghia Hoang, and Phi Le Nguyen. Federated prompt-tuning with heterogeneous and incomplete multi-modal client data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3936–3946, October 2025.

Atit Pokharel, Ratun Rahman, Thomas Morris, and Dinh C. Nguyen. Quantum federated learning for multimodal data: A modality-agnostic approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 545–554, June 2025.

Fan Qi and Shuai Li. Adaptive hyper-graph aggregation for modality-agnostic federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12312–12321, 2024.

Minhyuk Seo, Taeheon Kim, Hankook Lee, Jonghyun Choi, and Tinne Tuytelaars. Not all clients are equal: Collaborative model personalization on heterogeneous multi-modal clients, 2025. URL https://arxiv.org/abs/2506.11024.

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)*, pp. 247–263, 2018.

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*, pp. 12695–12705, 2020.

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024.

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*, pp. 24043–24055. PMLR, 2022.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3063–3071, 2022.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

Liangqi Yuan, Dong-Jun Han, Vishnu Pandi Chellapandi, Stanislaw H. Żak, and Christopher G. Brinton. Fedmfs: Federated multimodal fusion learning with selective modality communication, 2024. URL https://arxiv.org/abs/2310.07048.

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)*, pp. 2236–2246, 2018.

Huimin Zeng, Zhenrui Yue, and Dong Wang. Open-vocabulary federated learning with multimodal prototyping. *arXiv preprint arXiv:2404.01232*, 2024.

Han Zhang, Yiding Li, and Xuelong Li. Constrained bipartite graph learning for imbalanced multimodal retrieval. *IEEE Transactions on Multimedia*, 26:4502–4514, 2023a.

Rongyu Zhang, Xiaowei Chi, Guiliang Liu, Wenyi Zhang, Yuan Du, and Fangxin Wang. Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation. *arXiv preprint arXiv:2303.15486*, 2023b.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27456–27466, 2024.

Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pp. 1–15, 2023.

# A APPENDIX

Given the space limitations in the main paper, we defer additional materials to the appendix. Appendix A.1 provides further implementation and experimental setup details, Appendix A.2 reports additional experimental results, and Appendix A.3 contains the convergence analysis and corresponding proofs.

## A.1 ADDITIONAL EXPERIMENTAL SETUP

### A.1.1 DATASETS

**CREMA-D** CREMA-D is an audio-visual dataset for speech emotion recognition. It contains 7,442 original clips (2–3 seconds) from 91 actors speaking short utterances, annotated with six emotion classes (anger, disgust, fear, happy, neutral, and sad). Labels are obtained via crowd-sourcing from 2,443 raters.

**AVE** The Audio-Visual Event (AVE) dataset is a benchmark for audio-visual event localization, consisting of 4,143 YouTube videos from 28 event categories. Each video is temporally annotated with audio-visual event boundaries and each audio-visual event lasts at least 2 seconds. The events span diverse domains including human/animal activities, musical performances, and vehicle-related sounds.

**CMU-MOSEI** CMU-MOSEI is a large-scale multimodal dataset for sentiment and emotion analysis. In our experiments, we use 22,856 annotated utterances. Each utterance is associated with a sentiment score ranging from 3 (strongly negative) to +3 (strongly positive), which we map into three classes: negative, neutral, and positive.

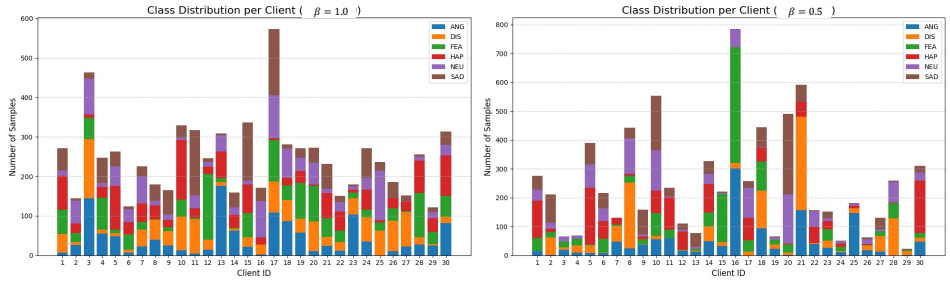
Table 4: Statistics of the datasets used in our experiments.

| Dataset | #sample | #class | #client | Modality | | |
|---|---|---|---|---|---|---|
| | | | | Visual | Audio | Text |
| CREMA-D | 7,442 | 6 | 30 | ✓ | ✓ | ✗ |
| AVE | 4,143 | 28 | 30 | ✓ | ✓ | ✗ |
| CMU-MOSEI | 22,856 | 3 | 70 | ✓ | ✓ | ✓ |

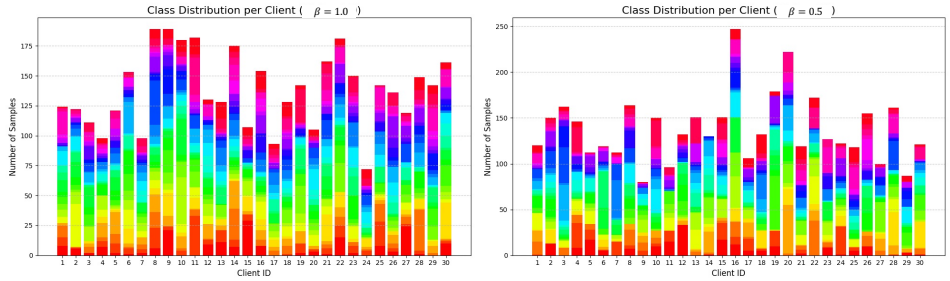
As shown in Figures 4 and 5, we present stacked bar charts of the client-wise label distributions under $\beta \in \{0.5, 1.0\}$. Each bar represents the number of samples assigned to a client, with colored segments indicating class-specific sample counts. In particular, a smaller concentration parameter ($\beta = 0.5$) leads to more skewed and client-specific label compositions, whereas a larger value ($\beta = 1.0$) yields comparatively more balanced label distributions across clients. As shown in Figure 6, we further illustrate the client-wise modality availability resulting from our modality partitioning strategy. Specifically, we construct a binary client–modality matrix in which each entry indicates whether a modality is available at a given client; the heatmap uses color to distinguish available versus missing modalities, with columns denoting clients and rows denoting modalities. Such modality-level non-IID settings reflect realistic multimodal federated scenarios, where acquisition or system constraints lead to systematic modality missingness across participants.

### A.1.2 TRAINING DETAILS

With respect to the CREMA-D and AVE datasets, we adopt a ResNet-18 backbone with modality-specific input stems (3-channel for vision and 1-channel for audio). For both datasets, we use one video frame of size $224 \times 224 \times 3$, extracted as the middle frame of each clip (with random resized cropping and horizontal flipping during training). For the audio stream, we load the waveform at $22{,}050$ Hz and ensure a fixed duration of 3 seconds by truncation or repetition. We then compute a log-magnitude STFT spectrogram using $\text{n_fft} = 512$ and $\text{hop_length} = 353$, resulting in $257$ frequency bins (i.e., $1 + \text{n_fft}/2$). The visual and audio backbones output $512$-dimensional global features via adaptive pooling, which are fed into a fully connected layer to produce modality-specific predictions. For multimodal inference, we adopt a simple late-fusion strategy by summing the




(a) CREMA-D




(b) AVE

Figure 4: Client-wise label distributions on CREMA-D and AVE datasets under different heterogeneity levels ($\beta \in \{0.5, 1.0\}$).

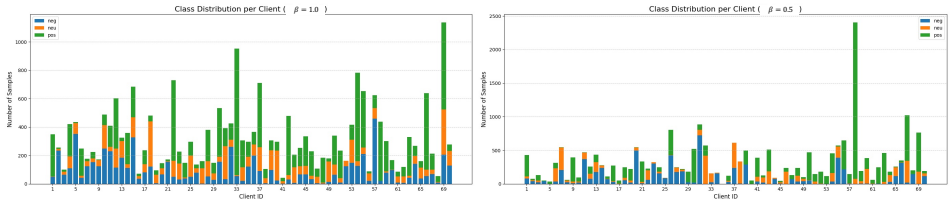



Figure 5: Client-wise label distributions on CMU-MOSEI under different heterogeneity levels ($\beta \in \{0.5, 1.0\}$).

modality-specific logits to obtain the final prediction. FEDMCHAIN itself is agnostic to the specific fusion operator and can accommodate alternative late-fusion schemes.

For the CMU-MOSEI dataset, we use pre-extracted language, visual, and acoustic features. The text modality is represented by BERT embeddings (Devlin et al., 2019) with a 768-dimensional feature size, while the visual and audio modalities use FACET features (35 dimensions) and COVAREP features (Degottex et al., 2014) (74 dimensions), respectively. These features are passed through modality-specific encoders to produce 128-dimensional latent representations. In particular, AudioNet and VisualNet adopt a three-layer MLP backbone, where each layer is followed by ReLU, dropout, and layer normalization, and then a linear layer is used to output modality-specific predictions. For the text modality, TextNet uses a bidirectional LSTM to encode the input sequence; the final hidden state is projected to a 128-dimensional vector and fed into a linear prediction head. We use the same logit-level late-fusion strategy to combine predictions from the available modalities.

For all three datasets, we use the Adam optimizer with dataset-specific learning rates of $1 \times 10^{-5}$, $1 \times 10^{-4}$, and $1 \times 10^{-3}$, respectively. In each training round, we randomly sample 70% of the clients to participate in training. All methods are trained for 500 rounds in total. For FEDMCHAIN, these rounds are distributed across the modality stages, with server aggregation performed every 20 training rounds; the baselines aggregate after every round. For the SSCA, we perform direction-consistency clustering using the standard KMeans algorithm.

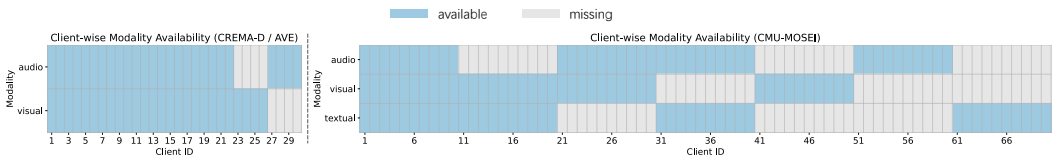
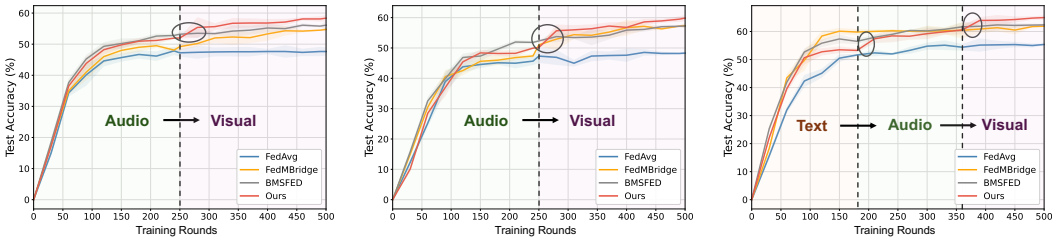



Figure 6: Client-wise modality availability on CREMA-D/AVE and CMU-MOSEI.




(a) CREMA-D (b) AVE (c) CMU-MOSEI

Figure 7: Performance over training rounds on the three datasets.

### A.2 Additional Experimental Results

**Training-Round Dynamics.** Figure 7 shows the average test performance across all clients over 500 training rounds. Due to the chained modality training paradigm, FedMChain exhibits a distinct phase-wise improvement pattern across modality stages. These improvements accumulate through successive modality updates, leading to stronger performance in later training rounds. Notably, all methods use the same number of training rounds, while the baselines aggregate after every round and FedMChain aggregates every 20 rounds, substantially reducing the frequency of server aggregation.

**Sensitivity of the Aggregation Period.** Table 5 compares different aggregation periods $\upsilon \in \{1, 5, 10, 15, 20, 30\}$, where the server aggregates once every $\upsilon$ training rounds. Performance generally improves as $\upsilon$ increases from 1 to 20, and then plateaus or slightly degrades for larger $\upsilon$. This trend suggests a trade-off between sufficient local optimization and client drift. A moderate aggregation period may provide more informative directional updates for SSCA, whereas excessively long local trajectories under non-IID data may increase client drift and weaken the benefit of consensus-based aggregation. Overall, FedMChain remains effective across a broad range of aggregation periods, with $\upsilon = 20$ providing the best overall trade-off in our experiments.

**Effect of the number of clusters.** We evaluate FedMChain with different numbers of clusters, setting $K \in \{2, 3, \ldots, 10\}$. Figure 8a reports test accuracy on CREMA-D, AVE, and CMU-MOSEI under $\beta \in \{0.5, 1.0\}$. Across datasets and both $\beta$ values, performance generally improves as $K$ increases from 2 to a moderate range and then plateaus or mildly degrades for larger $K$. The best or near-best performance is typically achieved around $K = 4 \sim 6$, suggesting that a moderate clustering granularity is sufficient to capture heterogeneous client-update patterns, while further partitioning provides limited additional benefit. Overall, the method is relatively insensitive within this moderate range, and we set $K = 5$ as the default in all experiments.

**Sensitivity of $\lambda_{\text{merge}}$ on the three datasets.** Figure 8b reports the effect of $\lambda_{\text{merge}}$ across the three datasets. Overall, performance peaks when $\lambda_{\text{merge}}$ lies in the range 0.8–0.9 (our final choice is $\lambda_{\text{merge}} = 0.9$). Since $\lambda_{\text{merge}}$ controls the contribution of the SSCA-aggregated consensus update to the global model, too small values render server updates overly conservative and slow to accumulate consensus. In contrast, too large values make the updates overly aggressive and more sensitive to residual cross-client inconsistency, slightly impairing stability.

**Sensitivity of hyperparameters $(\kappa, \pi)$ in SSCA.** Table 6 evaluates the hyperparameter sensitivity of SSCA by varying the retention ratio $\kappa$ and the consensus threshold $\pi$, two factors that directly control how much client update information is preserved and how conservatively cross-cluster updates are reconciled. Specifically, $\kappa$ retains the top-$\kappa$ fraction of coordinates (thus sparsifying $1 - \kappa$), while $\pi$ gates coordinate-wise cross-cluster merging according to the directional dominance criterion.

Table 5: Effect of the aggregation period on overall ACC (%) across the three datasets.

| *period* ($\upsilon$) | 1 | 5 | 10 | 15 | 20 | 30 |
|---|---|---|---|---|---|---|
| CREMA-D | 57.69 | 57.93 | 58.08 | 58.21 | **58.36** | 58.33 |
| AVE | 58.76 | 59.38 | 59.71 | 59.64 | **59.85** | 59.42 |
| CMU-MOSEI | 64.02 | 64.34 | 64.19 | 64.22 | **64.96** | 64.83 |

Table 6: Effect of $\kappa$ and $\pi$ on overall ACC (%) across the three datasets.

| $\kappa$ | $\kappa = 0.1$ | $\kappa = 0.2$ | $\kappa = 0.3$ | $\kappa = 0.4$ | $\kappa = 0.5$ | $\kappa = 0.6$ | $\kappa = 0.7$ | $\kappa = 0.8$ | $\kappa = 0.9$ |
|---|---|---|---|---|---|---|---|---|---|
| CREMA-D | 50.16 | 51.84 | 53.12 | 55.25 | 56.34 | 58.01 | **58.36** | 58.14 | 57.79 |
| AVE | 54.38 | 54.97 | 55.43 | 57.31 | 58.66 | 59.18 | **59.85** | 59.73 | 59.44 |
| CMU-MOSEI | 60.67 | 61.68 | 62.24 | 62.53 | 63.75 | 64.27 | 64.84 | **64.96** | 64.65 |
| $\pi$ | $\pi = 0.1$ | $\pi = 0.2$ | $\pi = 0.3$ | $\pi = 0.4$ | $\pi = 0.5$ | $\pi = 0.6$ | $\pi = 0.7$ | $\pi = 0.8$ | $\pi = 0.9$ |
| CREMA-D | 56.43 | 56.29 | 56.82 | 57.02 | 57.33 | 58.36 | 58.21 | **58.39** | 58.14 |
| AVE | 55.92 | 56.47 | 57.55 | 57.32 | 57.86 | 58.48 | 59.24 | **59.85** | 59.45 |
| CMU-MOSEI | 59.36 | 59.73 | 60.31 | 60.56 | 61.43 | 63.76 | 63.99 | **64.96** | 64.76 |

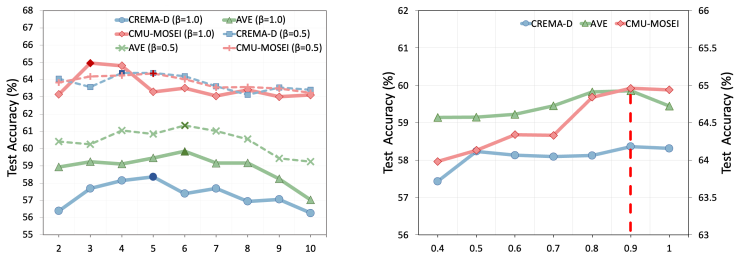
Regarding $\kappa$, performance improves markedly as $\kappa$ increases from 0.1 to the range 0.7–0.8, and then saturates or slightly degrades at larger values. The strongest performance is obtained at $\kappa = 0.7$ on CREMA-D and AVE and at $\kappa = 0.8$ on CMU-MOSEI, indicating that the method performs consistently well within this moderate range.

For $\pi$, moderate-to-high values generally outperform small thresholds, with the strongest overall performance observed around $\pi \approx 0.8$. This result is consistent with the role of $\pi$ in controlling cross-cluster merging: overly permissive or overly conservative merging can both limit the effectiveness of aggregation.

**Sensitivity of $\lambda_a$ and $\lambda_c$ on the three datasets.** Figure 9a–c shows the effect of $\lambda_a$ and $\lambda_c$ on CREMA-D, AVE, and CMU-MOSEI. Overall across the three datasets, the strongest performance is achieved when $\lambda_a$ is in the range $0.3$–$0.5$ and $\lambda_c$ is in the range $0.8$–$1.2$ (with our final choice $\lambda_a = 0.4$ and $\lambda_c = 1.0$ lying near the optimum). This observation suggests that moderate $\lambda_a$ provides sufficient cross-modal alignment to transfer stable semantics from preceding modalities, while moderate $\lambda_c$ strengthens complementary learning by emphasizing hard/uncertain samples; excessively small values weaken both effects, whereas overly large values over-regularize the stage objective and can amplify noisy reweighting signals, reducing generalization.

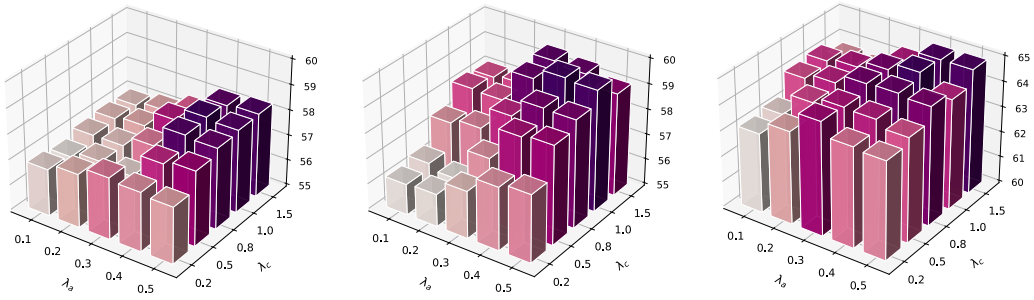
**Large-scale Client Scalability.** To evaluate FEDMCHAIN under a larger number of clients, we further conduct experiments on CMU-MOSEI with 200 clients. As shown in Table 7, FEDMCHAIN achieves the highest overall accuracy ($61.35\%$) and multimodal-client accuracy ($62.41\%$), outperforming BMSFed by $1.49$ and $1.44$ percentage points, respectively. It also obtains higher visual- and audio-modality accuracy, while remaining comparable on the text modality ($65.36\%$ vs. $65.42\%$). These results provide additional evidence that the performance gains of FEDMCHAIN persist when scaling the federated system to a substantially larger number of clients.

**Effect of Modality-wise Training-Round Allocation.** We further investigate whether different allocations of training rounds across modality stages affect FEDMCHAIN. In addition to the uniform allocation adopted in the main experiments, we consider three alternative strategies. Specifically, S1 and S2 use fixed non-uniform allocations, assigning training rounds to the audio and visual stages at ratios of 7:3 and 3:7, respectively, while S3 adaptively switches modality stages according to a training-loss plateau criterion. As shown in Table 8, under the same 500-round budget, S2 improves over uniform allocation by $0.32$ and $0.12$ percentage points on CREMA-D and AVE, respectively, whereas S1 yields lower accuracy on both datasets. Moreover, S3 achieves the highest accuracy while using only 462 and 484 training rounds, respectively. These results indicate that modality-wise round allocation can affect the final performance and that adaptive stage scheduling provides a promising complementary direction. We nevertheless use uniform allocation in the main experiments to isolate the effect of the proposed sequential optimization paradigm from additional scheduling strategies.




*(a) K* *(b) $\lambda_{\text{merge}}$*

Figure 8: (a) Effect of the number of SSCA clusters $K$ on test accuracy across three datasets under different data heterogeneity levels $\beta \in \{0.5, 1.0\}$. (b) Effect of $\lambda_{\text{merge}}$ on the performance of the three datasets.




*(a)* CREMA-D *(b)* AVE *(c)* CMU-MOSEI

Figure 9: The effect of $\lambda_a$ and $\lambda_c$ on the performance across the three datasets.

### A.2.1 Modality Imbalance Ratio (MIR)

We use the modality imbalance ratio (MIR) as a performance-based measure of modality imbalance associated with modality competition. For each modality $m$, we evaluate the model in a unimodal inference manner on all clients containing modality $m$, including both unimodal and multimodal clients, by enabling only modality $m$ and disabling/masking the other available modalities. We then average the resulting accuracies over these clients to obtain the modality-wise mean accuracy $\bar{\text{ACC}}^{(m)}$. Finally, MIR is defined as the ratio between the best and the worst modality-wise mean accuracies:

$$\text{MIR} = \frac{\max_m \ \bar{\text{ACC}}^{(m)}}{\min_m \ \bar{\text{ACC}}^{(m)}}. \tag{19}$$

By construction, $\text{MIR} \geq 1$, and a larger MIR indicates a larger modality-induced performance imbalance.

Table 7: Performance comparison on CMU-MOSEI with 200 clients. The best result in each column is highlighted in bold.

| Method | ACC | $\mathrm{ACC}_v$ | $\mathrm{ACC}_a$ | $\mathrm{ACC}_t$ | $\mathrm{ACC}_m$ |
|---|---|---|---|---|---|
| FedMBridge | 58.44 | 42.08 | 48.69 | 64.78 | 60.53 |
| BMSFed | 59.86 | 42.16 | 49.11 | **65.42** | 60.97 |
| FEDMCHAIN | **61.35** | **43.89** | **49.74** | 65.36 | **62.41** |

Table 8: Effect of different modality-wise training-round allocation strategies on CREMA-D and AVE. S1 and S2 allocate training rounds to the audio and visual stages at ratios of 7:3 and 3:7, respectively; S3 adaptively switches modality stages according to a training-loss plateau criterion; and Uniform denotes the equal allocation used in the main experiments.

| | CREMA-D | | AVE | |
|---|---|---|---|---|
| Strategy | ACC | Training Rounds | ACC | Training Rounds |
| S1 | 57.92 | 500 | 59.18 | 500 |
| S2 | 58.68 | 500 | 59.97 | 500 |
| S3 | **58.75** | 462 | **60.29** | 484 |
| Uniform (Ours) | 58.36 | 500 | 59.85 | 500 |

## A.3 Convergence Analysis

### A.3.1 Problem Setup

**Modality-stage view.** MCFT proceeds in modality-specific stages by activating one modality at a time. In a fixed stage $m$, only the parameter block associated with the active modality is updated, while all other modality branches are frozen (see main text). Accordingly, the analysis below focuses on the stage-$m$ parameter block, denoted by $\Theta \in \mathbb{R}^{d_m}$.

**Cluster-specific objective.** For analysis, we condition on a fixed cluster partition $\{\mathcal{C}_k\}_{k=1}^{K}$ within modality stage $m$. Fix $k \in \{1, \dots, K\}$ and let $\mathcal{C}_k$ denote the corresponding set of clients. We define the stage-$m$, cluster-$k$ objective as

$$F^{(m,k)}(\Theta) := \sum_{i \in \mathcal{C}_k} p_i \, \mathbb{E}_{\xi \sim \mathcal{D}_i} \left[ f_i^{(m)}(\Theta; \xi) \right], \quad p_i \geq 0, \ \sum_{i \in \mathcal{C}_k} p_i = 1, \tag{20}$$

where $p_i = w_i / \sum_{j \in \mathcal{C}_k} w_j$ is the normalized client weight within cluster $k$, and $f_i^{(m)}(\Theta; \xi)$ denotes the stochastic stage-$m$ objective corresponding to the client loss $\mathcal{L}_i^{(m)}$ in Eq. (4). For brevity, throughout the remainder of Appendix A.3, we write $F(\Theta) \equiv F^{(m,k)}(\Theta)$ for the fixed stage-$m$, cluster-$k$ objective.

**Local update and SSCA aggregation.** Let $r = 0, 1, 2, \dots$ index the server aggregation rounds for the cluster model $k$ in stage $m$. At the beginning of round $r$, the server broadcasts $\Theta_r \equiv \Theta_r^{(m,k)}$ to clients in $\mathcal{C}_k$. Each client $i \in \mathcal{C}_k$ performs $\mathcal{E}$ local SGD steps:

$$\Theta_{i,r}^{\mathrm{s}+1} = \Theta_{i,r}^{\mathrm{s}} - \eta \, g_i^{(m)}(\Theta_{i,r}^{\mathrm{s}}; \xi_{i,r}^{\mathrm{s}}), \ \mathrm{s} = 0, 1, \dots, \mathcal{E} - 1, \ \Theta_{\mathrm{i,r}}^{0} = \Theta_{\mathrm{r}}, \tag{21}$$

where $g_i^{(m)}(\cdot; \xi)$ is a stochastic gradient of $f_i^{(m)}$ at sample $\xi$ and $\eta > 0$ is the local stepsize.

Define the client vector:

$$\Delta\Theta_{i,r} := \Theta_{i,r}^{\mathcal{E}} - \Theta_r. \tag{22}$$

SSCA takes $\{\Delta\Theta_{i,r}\}_{i \in \mathcal{C}_k}$ as input and outputs an aggregated update direction for cluster $k$:

$$\Delta\widehat{\Theta}_r \equiv \Delta\widehat{\Theta}_r^{(m,k)} \in \mathbb{R}^{d_m}. \tag{23}$$

The server updates the cluster-$k$ model with merge strength $\lambda_{\mathrm{merge}} > 0$:

$$\Theta_{r+1} = \Theta_r + \lambda_{\mathrm{merge}} \, \eta_g \Delta\widehat{\Theta}_r. \tag{24}$$

**Notation.** The server update in the main text uses a single coefficient $\lambda_{\text{merge}}$(main). For analysis, we rewrite it as $\lambda_{\text{merge}}(\text{main}) = \eta_{\text{g}}\lambda_{\text{merge}}$, where $\eta_{\text{g}}$ is a global stepsize and $\lambda_{\text{merge}}$ denotes the merge strength. This is a purely notational re-parameterization and does not change the algorithm.

### A.3.2 Assumptions

**Assumption A.1** (Smoothness)**.** The fixed stage-$m$, cluster-$k$ objective $F(\Theta)$ is $L$-smooth, i.e., for all $\Theta, \Theta'$, $||\nabla F(\Theta) - \nabla F(\Theta')|| \le L||\Theta - \Theta'||$.

**Assumption A.2** (Expected descent correlation of SSCA update)**.** There exists a constant $\gamma > 0$ such that for all rounds $r$,

$$\Big\langle \nabla F(\Theta_r),\ \mathbb{E}\big[\Delta\widehat{\Theta}_r \mid \Theta_r\big] \Big\rangle \ \le\ -\gamma \left\|\nabla F(\Theta_r)\right\|^2. \tag{25}$$

**Assumption A.3** (Bounded second moment of SSCA update)**.** Assume that $\mathbb{E}\Big[\|\Delta\widehat{\Theta}_r\|^2 \mid \Theta_r\Big]$ exists and is finite for all rounds $r$. There exist constants $B > 0$ and $\sigma^2_{\text{agg}} \ge 0$ such that for all rounds $r$,

$$\mathbb{E}\big[\|\Delta\widehat{\Theta}_r\|^2 \mid \Theta_r\big] \ \le\ B\,\|\nabla F(\Theta_r)\|^2 + \sigma^2_{\text{agg}}. \tag{26}$$

*Remark* A.4 (Interpretation of Assumptions A.2–A.3). Assumption A.2 requires the expected SSCA update to be descent-correlated with the cluster objective whenever $\nabla F(\Theta_r) \neq 0$. Assumption A.3 bounds the conditional second moment of the aggregated update and permits a residual term $\sigma^2_{\text{agg}}$ that is independent of the current gradient norm.

### A.3.3 SSCA as an Inexact Descent Direction

**From local deltas to SSCA direction.** SSCA maps the collection of local deltas $\{\Delta\Theta_{i,r}\}_{i\in\mathcal{C}_k}$ to $\Delta\widehat{\Theta}_r$. We model SSCA as an aggregation operator that returns a descent-correlated update with a bounded conditional second moment, formalized in Assumptions A.2 and A.3.

### A.3.4 Main result: Nonconvex convergence to a neighborhood

**Theorem A.5** (Stage-$m$, cluster-$k$ nonconvex convergence of FedMChain)**.** *Suppose Assumption A.1 holds and $F^* := \inf_\Theta F(\Theta) > -\infty$. Assume further that Assumptions A.2–A.3 hold for the* actual *SSCA direction $\Delta\widehat{\Theta}_r$ produced by SSCA at round $r$. Choose the* server *stepsize $\eta_{\text{g}} > 0$ such that*

$$\gamma - \frac{L\,\eta_{\text{g}}\,\lambda_{\text{merge}}\,B}{2} \ >\ 0 \qquad \textit{(e.g., } \eta_{\text{g}} < \frac{2\gamma}{L\lambda_{\text{merge}}B} \textit{ when } B > 0\textit{)}. \tag{27}$$

*Then for any number of aggregation rounds $R \ge 1$,*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla F(\Theta_r)\right\|^2 \ \le\ \frac{2\big(F(\Theta_0) - F^*\big)}{\eta_{\text{g}}\,\lambda_{\text{merge}}\Big(2\gamma - L\eta_{\text{g}}\lambda_{\text{merge}}B\Big)\,R} \ +\ \frac{L\,\eta_{\text{g}}\,\lambda_{\text{merge}}\,\sigma^2_{\text{agg}}}{\Big(2\gamma - L\eta_{\text{g}}\lambda_{\text{merge}}B\Big)}, \tag{28}$$

*Proof.* By $L$-smoothness (Assumption A.1) and the update Eq. (24),

$$\begin{aligned} F(\Theta_{r+1}) &\le F(\Theta_r) + \big\langle \nabla F(\Theta_r),\ \Theta_{r+1} - \Theta_r\big\rangle + \frac{L}{2}\left\|\Theta_{r+1} - \Theta_r\right\|^2 \\ &= F(\Theta_r) + \eta_{\text{g}}\lambda_{\text{merge}}\big\langle \nabla F(\Theta_r),\ \Delta\widehat{\Theta}_r\big\rangle + \frac{L\eta_{\text{g}}^2\lambda_{\text{merge}}^2}{2}\left\|\Delta\widehat{\Theta}_r\right\|^2. \end{aligned} \tag{29}$$

Take conditional expectation w.r.t. $\Theta_r$ and apply Assumptions A.2– A.3:

$$\begin{aligned} \mathbb{E}\big[F(\Theta_{r+1}) \mid \Theta_r\big] &\le F(\Theta_r) - \eta_{\text{g}}\lambda_{\text{merge}}\gamma\|\nabla F(\Theta_r)\|^2 + \frac{L\eta_{\text{g}}^2\lambda_{\text{merge}}^2}{2}\Big(B\|\nabla F(\Theta_r)\|^2 + \sigma^2_{\text{agg}}\Big) \\ &= F(\Theta_r) - \eta_{\text{g}}\lambda_{\text{merge}}\Big(\gamma - \frac{L\eta_{\text{g}}\lambda_{\text{merge}}B}{2}\Big)\|\nabla F(\Theta_r)\|^2 + \frac{L\eta_{\text{g}}^2\lambda_{\text{merge}}^2}{2}\sigma^2_{\text{agg}}. \end{aligned} \tag{30}$$

Now take total expectation and sum Eq. (30) over $r = 0, \ldots, R-1$:

$$\eta_{\mathrm{g}}\lambda_{\mathrm{merge}}\Big(\gamma - \frac{L\eta_{\mathrm{g}}\lambda_{\mathrm{merge}}B}{2}\Big)\sum_{r=0}^{R-1}\mathbb{E}\|\nabla F(\Theta_r)\|^2 \le F(\Theta_0) - \mathbb{E}F(\Theta_R) + \frac{L\eta_{\mathrm{g}}^2\lambda_{\mathrm{merge}}^2}{2}R\sigma_{\mathrm{agg}}^2. \quad (31)$$

Using $F(\Theta_R) \ge F^*$ and dividing both sides by $\eta_{\mathrm{g}}\lambda_{\mathrm{merge}}\big(\gamma - \frac{L\eta_{\mathrm{g}}\lambda_{\mathrm{merge}}B}{2}\big)R$ yields Eq.(28). □

*Remark* A.6 (Choosing stepsizes). Condition Eq.(27) ensures a *positive descent coefficient*. When $B > 0$, it is sufficient to choose $\eta_{\mathrm{g}} < \frac{2\gamma}{L\lambda_{\mathrm{merge}}B}$.