MULTIMODAL FEDERATED LEARNING WITH MODALITY-IMPORTANCE-AWARE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Federated Learning (MFL) has drawn significant attention as a decentralized training paradigm across multimodal edge devices. However, its communication cost, which grows linearly with the number of modalities, has emerged as a distinct scalability challenge. While prior approaches mitigate this through selective transmission of modality updates, they may sacrifice complementary information essential for multimodal learning. In this paper, we introduce a modality-importance-aware compression framework for MFL, termed `Fed-MIAC`, which retains multimodal benefits while improving communication efficiency. Specifically, `Fed-MIAC` differentiates compression rates across modality-specific model updates by jointly accounting for their relative importance and compression-induced distortion. Extensive simulations on diverse MFL benchmarks demonstrate that `Fed-MIAC` better preserves multimodal performance than importance-agnostic compression and modality selection baselines under constrained communication budgets, also validated by wall-clock overhead evaluations on embedded edge platforms. We further provide a convergence analysis that characterizes how modality-importance-aware compression affects the convergence rate.

1 INTRODUCTION

With the rapid proliferation of edge devices, such as Internet of Things nodes and wearable devices, large volumes of user data are generated and stored locally, creating opportunities for deep learning-based applications. However, the conventional centralized learning paradigm requires aggregating local data at a central server, which is especially undesirable when the data contains privacy-sensitive information such as health records or personal data. Federated Learning (FL) addresses this issue by enabling distributed training of a global model using local samples without sharing raw data.

Alongside its privacy advantages, FL introduces operational challenges, most notably communication overhead. Particularly in practical cross-device FL deployments, clients must repeatedly transmit their model updates over bandwidth-limited wireless links until convergence, imposing a substantial burden on resource-constrained edge devices. Accordingly, a large body of work has explored compression schemes to reduce uplink payloads and enable communication-efficient FL (Reisizadeh et al., 2020; Wang et al., 2022a; 2023; Yang et al., 2025). However, most existing methods are designed and evaluated under unimodal settings, leaving a significant gap with emerging on-device multimodal applications, such as wearable-based activity recognition, mobile image-text models, and healthcare monitoring (Zhang et al., 2022; Vasu et al., 2024; Zuo et al., 2026).

This gap calls into question the suitability of directly extending unimodal FL and applying modality-agnostic compression methods to multimodal FL (MFL). In particular, multimodal learning combines modality-specific pathways with cross-modal fusion to exploit complementary information and learn richer joint representations, while transmitting the resulting modality-wise model updates incurs substantial communication overhead that scales linearly with the number of modalities.

The fundamental challenge in MFL is therefore to achieve communication efficiency while minimally sacrificing the benefits of multimodal learning. However, existing methods do not fully address this trade-off. Specifically, recent modality selection-based approaches drop a subset of modality updates from transmission to mitigate the scalability challenge, but may consequently discard complementary information and weaken cross-modal synergy (Li et al., 2024; Yuan et al., 2026). This limitation motivates an alternative to prior binary modality selection approaches.

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 this end, we propose Fed-MIAC, a framework that integrates modality importance into model compression for communication-efficient MFL. The key observation underlying Fed-MIAC is that the contribution of each modality to the target task, i.e., its relative importance, can vary across clients and communication rounds. This naturally motivates each client to differentiate compression rates across modality-specific components based on locally estimated modality importance. To operationalize this idea, we formulate modality-importance-aware compression as an optimization problem that minimizes importance-weighted distortion under a shared communication budget. We then instantiate this formulation with singular-value-decomposition (SVD)-based compression and derive an optimal modality-wise rank-allocation rule. Our contributions are summarized as follows:

- We introduce Fed-MIAC, a modality-importance-aware compression framework for communication-efficient MFL that moves beyond a binary modality selection.
- We instantiate Fed-MIAC with SVD-based compression and derive an optimal rank-allocation rule that minimizes importance-weighted distortion under a shared rank budget.
- We extensively evaluate Fed-MIAC on CrisisMMD, KU-HAR, PTB-XL, and CMU-MOSEI benchmarks, demonstrating consistent performance gains over existing communication-efficient MFL baselines across a wide range of communication budgets. We further validate its practicality through a real-device overhead analysis on NVIDIA Jetson TX2, accounting for both computation and communication costs.
- We provide a convergence analysis that characterizes how modality-wise compression residuals affect MFL optimization and identifies compression-dependent terms in the bound.

2 Related Works

Communication-efficient federated learning: Communication efficiency is a critical systems requirement in cross-device FL deployments, as repeated transmission of model updates over capacity-limited wireless links can impose a substantial burden on edge devices and increase communication-round latency (Reisizadeh et al., 2020). More importantly, such a communication bottleneck can directly affect the distributed learning process: transmission outages may impair convergence (Wang et al., 2022a), while slow client uploads can create stragglers that delay synchronous aggregation.

A prominent line of work addresses these constraints by adapting established compression techniques to efficiently represent model parameters and updates for transmission. Specifically, FedPAQ (Reisizadeh et al., 2020) combines periodic averaging with quantized local model differences, while FedMPQ (Yang et al., 2025) employs multi-codebook product quantization to compress client updates. Another line of work exploits low-rank structure through SVD-based representations. FedDLR (Qiao et al., 2021) applies truncated SVD with spectral-energy-based rank selection, whereas SVDFed (Wang et al., 2023) exploits gradient correlations across clients and communication rounds by reusing shared SVD bases and transmitting client-specific coefficients between basis updates.

Multimodal federated learning (MFL): Multimodal learning exploits complementary information through representation learning and fusion across multiple modalities (Baltrušaitis et al., 2019). By revisiting both local learning mechanisms and communication strategies across clients, recent studies have sought to realize these benefits in the distributed learning paradigm. On the learning side, CreamFL (Yu et al., 2023) employs contrastive learning to align modality-specific representations and encourage cross-modal consistency, while MFCPL (Le et al., 2025) guides local representation learning through prototype-based regularization across modalities and classes. On the communication side, prior studies reduce the communication overhead associated with multiple modality-specific encoders by selectively transmitting their updates. Specifically, MCB (Li et al., 2024) selects client–modality pairs by balancing modality contributions against model sizes, while MFedMC (Yuan et al., 2026) additionally considers upload recency to select modality updates.

Focusing on the communication side, we note that modality selection approaches can be interpreted within a broader resource allocation framework across modalities, where the total communication budget is constrained by available uplink throughput and the delay allowed per round. From this perspective, binary modality selection corresponds to assigning either zero or full rate to each update, but this does not necessarily yield an optimal allocation when intermediate rate options are available. To address this, Fed-MIAC differentiates compression levels across modalities and formulates rate allocation as importance-weighted distortion minimization under a shared communication budget.

3 PRELIMINARIES AND PROBLEM FORMULATION

In this section, we review preliminaries of MFL and formulate the modality-importance-aware compression problem. We first introduce FedAvg (McMahan et al., 2017) as the FL baseline and extend the notation to the multimodal setting. Next, we review truncated SVD as a compression operator with a tractable rate–distortion relationship. Finally, we formulate the optimization problem by considering both modality importance and the distortion induced by rank allocation.

3.1 PRELIMINARIES

Federated learning (FL): We consider a FL system with one server and N clients. Client $i \in [N]$ holds n_i local samples, and we denote the total number of samples by $n := \sum_{i=1}^{N} n_i$, where $[N]$ is a set of integers between 1 and N. A data sample $z_{i,j} = (x_{i,j}, y_{i,j})$ is the j-th sample at client i with a paired input $x_{i,j}$ and its label $y_{i,j}$, and it is drawn from the local data distribution $\mathcal{D}_i$. The local objective function of client i is $F_i(w) := \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(w; z_{i,j})$, where w is the model weight and $\ell(\cdot, \cdot)$ is the loss function. Accordingly, FL aims to minimize the following global objective:

$$F(w) := \sum_{i=1}^{N} \frac{n_i}{n} F_i(w). \tag{1}$$

To minimize Eq. (1), we adopt FedAvg (McMahan et al., 2017) as a standard FL baseline. At round r, the server broadcasts the current global model w^r to a subset of active clients $\mathcal{S}^r \subset [N]$. Each selected client initializes its local training with $w_{i,0}^r = w^r$ and performs T steps of local SGD:

$$w_{i,t+1}^r = w_{i,t}^r - \eta \nabla \ell\big(w_{i,t}^r; \xi_{i,t}\big), \quad t = 0, \ldots, T-1,$$

where η is the local learning rate and $\xi_{i,t}$ is a mini-batch sampled from client i's local dataset. After local training, client i uploads the update $\Delta_i^r := w_{i,T}^r - w^r$, i.e., the difference between locally trained and global models. The server then aggregates the received updates by a weighted average:

$$w^{r+1} = \sum_{i \in \mathcal{S}^r} \frac{n_i}{\sum_{j \in \mathcal{S}^r} n_j} w_{i,T}^r = w^r + \sum_{i \in \mathcal{S}^r} \frac{n_i}{\sum_{j \in \mathcal{S}^r} n_j} \Delta_i^r.$$

This procedure is repeated for R rounds.

Multimodal federated learning (MFL): MFL is an extension of FL, where each local sample is multimodal. In an M-modality setting, the j-th sample at client i is denoted by $z_{i,j} = (x_{i,j}^1, x_{i,j}^2, \ldots, x_{i,j}^M, y_{i,j})$, where $x_{i,j}^m$ is the input of modality m and $y_{i,j}$ is the label. A multimodal model consists of modality-specific encoders $\{h_m\}_{m=1}^M$, a fusion module h, which is a concatenation or an attention-based module, and a classifier g. Given $z_{i,j}$, each encoder produces a modality representation $h_m(x_{i,j}^m)$, which is fused by h into a joint feature and mapped to the prediction by g. Notably, modality-specific encoders h_m contain many more parameters than a fusion module h and a classifier g, so our formulations in the following section focus on modality-specific weights.

Singular value decomposition (SVD): We next summarize SVD as a compression primitive for controlling the communication cost of weight matrix transmissions. For any matrix $W \in \mathbb{R}^{p \times q}$, the rank-$\Lambda$ truncated SVD approximation is $W^{(\Lambda)} \approx U_\Lambda \Sigma_\Lambda V_\Lambda^\top$, where $U_\Lambda \in \mathbb{R}^{p \times \Lambda}$, $\Sigma_\Lambda \in \mathbb{R}^{\Lambda \times \Lambda}$ is diagonal, and $V_\Lambda^\top \in \mathbb{R}^{\Lambda \times q}$. Instead of transmitting all pq entries of W, we transmit U_Λ, the Λ diagonal entries of Σ_Λ, and $V_\Lambda^\top$, requiring $p\Lambda + \Lambda + \Lambda q = \Lambda(p+q+1)$ parameters. SVD can reduce the number of communicated bits compared to full-matrix transmission when $\Lambda(p+q+1) < pq$.

Modality importance: Let ϕ_m denote the importance score of modality m, for each $m \in [M]$. To capture their relative importance, we normalize the score via temperature-scaled softmax:

$$\alpha_m = \frac{\exp(\phi_m/\tau)}{\sum_{k=1}^{M} \exp(\phi_k/\tau)}, \qquad \tau > 0, \tag{2}$$

where τ controls the sharpness of the weighting distribution. A smaller τ yields a peaked distribution that emphasizes the most important modalities, while a larger τ produces a flatter weighting.

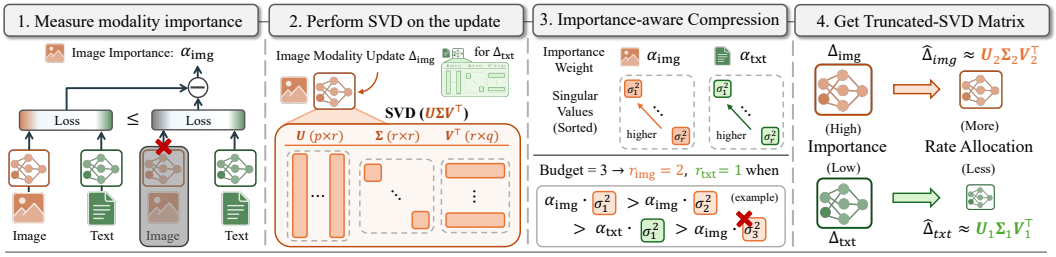

Figure 1: Proposed algorithm: Modality-importance-aware compression process of Fed-MIAC

3.2 Problem Formulation

We formulate modality-importance-aware compression as an optimization problem that minimizes importance-weighted distortion under a shared communication budget. For simplicity, we omit the client and round indices and denote the update matrix for modality m by $W_m \in \mathbb{R}^{p_m \times q_m}$. Here, we instantiate this operation with truncated SVD, where the compression level of modality m is controlled by its retained rank Λ_m. To control the overall communication cost, we further impose a total rank budget Λ shared across modalities, requiring $\sum_{m=1}^{M} \Lambda_m \leq \Lambda$.

Distortion from truncated SVD: For a matrix W and its rank-Λ truncated SVD approximation $W^{(\Lambda)}$, we define the distortion as the squared Frobenius norm of the residual:

$$D(W;\Lambda) \triangleq \left\|W - W^{(\Lambda)}\right\|_F^2 = \sum_{i>\Lambda} \sigma_i(W)^2, \tag{3}$$

where $\{\sigma_i(W)\}$ are the singular values of W in non-increasing order. The squared Frobenius norm measures the total squared energy of a matrix and admits a closed-form spectral characterization, making it well-suited for quantifying the overall distortion induced by truncated SVD.

Modality-importance-aware rank allocation problem: We formulate the rank allocation problem:

$$\min_{\{\Lambda_m\}_{m=1}^M} \sum_{m=1}^{M} \alpha_m D(W_m;\Lambda_m) \quad \text{s.t.} \quad \sum_{m=1}^{M} \Lambda_m \leq \Lambda, \quad \Lambda_m \in \mathbb{Z}_{\geq 0}. \tag{4}$$

This formulation captures a principled trade-off between communication and update fidelity across modalities. Specifically, Λ_m controls the fidelity of the transmitted update for modality m: a larger Λ_m reduces the truncation error at the cost of consuming more of the total rank budget Λ.

4 Proposed Method: Fed-MIAC

Building on the above formulation, we describe how Fed-MIAC combines local modality-importance estimation with optimal rank allocation to compress modality-specific updates. Given that client-side power constraints make uplink transmission the dominant bottleneck in FL deployments, we mainly focus on compressing the client updates (Reisizadeh et al., 2020).

Algorithm overview: In communication round r, the server samples a subset of clients $\mathcal{S}^r$ and broadcasts the current global model w^r (without compression). Each selected client $i \in \mathcal{S}^r$ performs T local training steps starting from w^r and obtains a local update $\Delta_i^r = w_{i,T}^r - w^r$. Subsequently, we partition the local update Δ_i^r into modality-specific blocks $\{\Delta_{i,m}^r\}_{m=1}^M$ (e.g., modality encoders), where m is a modality index. Fed-MIAC compresses them by allocating modality-wise ranks Λ_m under a fixed total rank budget Λ based on modality importance.

Compression process of Fed-MIAC: We illustrate the compression pipeline of Fed-MIAC in Figure 1 and the overall FL procedure, including the compression pipeline, is summarized in Appendix A. As shown in Figure 1, the four-step compression process is performed on the client side in each communication round. Details of these four steps are described as follows:

(Step 1) Measure modality importance: Shapley value (Shapley, 1953) provides a principled measure of the expected marginal contribution of each factor to model utility across coalitions.

In the multimodal setting, each modality is treated as a factor (or player) in a coalition game, and its contribution is quantified by the SHAPE score[1] (Hu et al., 2022), i.e., a modality-wise Shapley-based importance score. Let $[M] = \{1, \dots, M\}$ be the set of modalities, and let $M_S \subseteq [M]$ denote a subset of modalities. For client i, let $u_i(M_S)$ denote the utility (e.g., the negative loss) of using modalities in M_S. The SHAPE score of modality m at client i and round r is then defined as

$$\phi_{i,m}^r \triangleq \sum_{M_S \subseteq [M] \setminus \{m\}} \frac{|M_S|!(M - |M_S| - 1)!}{M!} \Big(u_i(M_S \cup \{m\}) - u_i(M_S) \Big),$$

where $|\cdot|$ denotes the cardinality of the set. We additionally normalize $\{\phi_{i,m}^r\}_{m=1}^M$ into modality-importance weights $\{\alpha_{i,m}^r\}_{m=1}^M$ via Softmax (referring to Eq. (2)).

(Step 2) Perform SVD on the update: After local training, each client computes its singular values with SVD for each modality-specific update block $\Delta_{i,m}^r$:

$$\Delta_{i,m}^r = U_{i,m}^r \Sigma_{i,m}^r (V_{i,m}^r)^\top.$$

Singular values in $\Sigma_{i,m}^r$ quantify distortion reduction when keeping additional rank in each modality.

(Step 3) Importance-aware compression: At client i and round r, we instantiate the problem formulation in Eq. (4) with $W_m = \Delta_{i,m}^r$, $\alpha_m = \alpha_{i,m}^r$, and decision variables $\Lambda_m = \Lambda_{i,m}^r$ under $\sum_{m=1}^M \Lambda_{i,m}^r \le \Lambda$. This yields the following optimal allocation rule.

Proposition 4.1 (Optimal modality-importance-aware rank allocation)**.** *Let the singular values of W_m be $\{\sigma_k(W_m)\}_{k=1}^{d_m}$ in non-increasing order, where $d_m = \min\{p_m, q_m\}$, and define the weighted marginal gain as $s_{m,k} := \alpha_m \sigma_k(W_m)^2$. Let $s_{(1)} \ge s_{(2)} \ge \cdots$ be the ordered values of $\{s_{m,k}\}_{m,k}$. Assume that the Λ-th largest weighted marginal gain is strictly greater than the next largest gain, i.e., $s_{(\Lambda)} > s_{(\Lambda+1)}$. Then the unique optimal solution to Eq. (4) is obtained by selecting the Λ modality–rank pairs corresponding to $s_{(1)}, \dots, s_{(\Lambda)}$. The allocated rank for modality m is*

$$\Lambda_m^\star \;=\; \big|\{k : \; (m,k) \textit{ is selected}\}\big|.$$

Remark 1. (Allocation rule) From Eq. (3), adding one rank to modality m (from $k-1$ to k) reduces the distortion by

$$D(W_m; k-1) - D(W_m; k) = \sigma_k(W_m)^2, \tag{5}$$

so the objective decreases by the weighted marginal gain $\alpha_m \sigma_k(W_m)^2$. Therefore, not only the modality importance, which is a property of the MFL, but also the distortion introduced by each model update affects the allocation rule. We provide the detailed proof in Appendix B.

(Step 4) Get truncated-SVD factors and upload: Using the allocation $\{\Lambda_{i,m}^r\}$, which is obtained at Step 3, client i forms the truncated SVD factors for each modality[2]:

$$U_{i,m}^r(:, 1{:}\Lambda_{i,m}^r),\ \Sigma_{i,m}^r(1{:}\Lambda_{i,m}^r, 1{:}\Lambda_{i,m}^r),\ V_{i,m}^r(:, 1{:}\Lambda_{i,m}^r),$$

and uploads these low-rank factors to the server. Thus, the number of elements to transmit decreases from Mpq to $\sum_{m=1}^M \Lambda_{i,m}^r(p+q+1)$, where $W_m \in \mathbb{R}^{p \times q}$. After the server receives the parameters with truncated SVD, the server reconstructs the compressed modality updates, i.e.,

$$\widehat{\Delta}_{i,m}^r = U_{i,m}^r(:, 1{:}\Lambda_{i,m}^r)\, \Sigma_{i,m}^r(1{:}\Lambda_{i,m}^r, 1{:}\Lambda_{i,m}^r)\, \big(V_{i,m}^r(:, 1{:}\Lambda_{i,m}^r)\big)^\top,$$

and concatenates them into $\widehat{\Delta}_i^r$ via aggregation across clients:

$$\widehat{\Delta}^r = \sum_{i \in \mathcal{S}^r} \frac{n_i}{\sum_{k \in \mathcal{S}^r} n_k} \widehat{\Delta}_i^r, \quad w^{r+1} = w^r + \widehat{\Delta}^r.$$

[1] We provide a discussion of Shapley value in the Appendix D.5.

[2] For a 2D matrix A, $A(:, i{:}j)$ means a sub-matrix with full row and partial columns indexing from i to j.

5 THEORETICAL ANALYSIS

Before presenting the theoretical analysis of `Fed-MIAC`, we introduce Assumptions C.1–C.4, which are standard in non-convex FL convergence analyses (McMahan et al., 2017; Wang et al., 2022b) and characterize the smoothness of the local objectives, the variance of stochastic gradients, bounded gradient divergence, and the boundedness of stochastic gradient norms, respectively.

Assumption C.5 imposes a uniform bound on the distortion induced by compression. For truncated SVD, the compression-error norm is bounded by the norm of the corresponding uncompressed update, with equality when the retained rank is zero. These assumptions are formally stated in Appendix C. Under these assumptions, we analyze the convergence of MFL under model compression.

5.1 CONVERGENCE ANALYSIS OF COMPRESSION MFL

We present a convergence analysis for SVD-compressed MFL that characterizes how modality-wise compression residuals affect optimization. In particular, our analysis characterizes the effect of the parameter-space compression residual $\eta e_{i,m}^r := \widehat{\Delta}_{i,m}^r - \Delta_{i,m}^r$, where the learning rate η is factored out because both the local model update and its corresponding SVD compression residual scale with the local learning rate. We provide a detailed proof in Appendix C.

Theorem 5.1. *Under Assumptions C.1–C.5, full participation, $\eta = 1/\sqrt{R}$, and $L\eta T \le 1/8$, the averaged squared gradient norm of SVD-compressed multimodal FL is upper bounded by:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \le \mathcal{B}_{\mathrm{FL}} + \frac{L}{TR^{3/2}}\sum_{r=0}^{R-1}\left(4\mathcal{M}_{\mathrm{dir}}^r + 2\mathcal{E}_{\mathrm{agg}}^r\right) + \frac{4}{TR}\sum_{r=0}^{R-1}\left(\mathcal{D}_{\mathrm{gap}}^r\right)^{1/2},$$

where $\mathcal{B}_{\mathrm{FL}}$ collects the standard federated optimization error terms. The terms caused by compression are $\mathcal{D}_{\mathrm{gap}}^r$, $\mathcal{M}_{\mathrm{dir}}^r$, and $\mathcal{E}_{\mathrm{agg}}^r$, denoting the gradient-weighted compression distortion, the aggregated local direction with distortion, and aggregated compression error energy, respectively. Their explicit definitions and the terms from which they are derived are given in Appendix C.7.

Remark 2. (Baseline term) The first term, $\mathcal{B}_{\mathrm{FL}}$, collects the standard optimization and heterogeneity terms that appear in the convergence bound under Assumptions C.1–C.4, where $\mathcal{B}_{\mathrm{FL}}$ behaves as $\mathcal{O}(1/\sqrt{R})$ in convergence. For the generality of our claim, when setting $e_i^r \leftarrow 0$ (no compression), Theorem 5.1 exhibits the convergence rate of FedAvg, since $\mathcal{D}_{\mathrm{gap}}^r$, $\mathcal{M}_{\mathrm{dir}}^r$, and $\mathcal{E}_{\mathrm{agg}}^r$ reduce to zero.

Remark 3. (A potential convergence gap under compression) The compression error-dependent part of the bound consists of two components. The second term in Theorem 5.1 decays as $\mathcal{O}(1/\sqrt{R})$, since $\mathcal{M}_{\mathrm{dir}}^r$ and $\mathcal{E}_{\mathrm{agg}}^r$ are uniformly bounded under the stated assumptions. In contrast, the last term is not guaranteed to vanish as $R \to \infty$, remaining a potential convergence gap.

$$\mathcal{D}_{\mathrm{gap}}^r := \mathbb{E}\left[G^2\sum_{i=1}^{N}\sum_{m=1}^{M} p_i \cdot \frac{\|\nabla_{w_m} F(w^r)\|_F}{G} \cdot \|e_{i,m}^r\|_F^2\right],$$

where $G := \sum_{k=1}^{M}\|\nabla_{w_k} F(w^r)\|_F$. This quantity arises from bounding the interaction between the global gradient and the aggregated compression residual, i.e., $\langle \nabla F(w^r), \sum_{i=1}^{N} p_i e_i^r\rangle$. Nevertheless, the compression-error bound $\|e_{i,m}^r\|_F^2 \le e^2$, together with the assumed gradient bound, ensures that both $\mathcal{D}_{\mathrm{gap}}^r$ and its contribution to the convergence bound remain uniformly bounded.

Although the gradient-based weights $\|\nabla_{w_m} F(w^r)\|_F / G$ are unavailable locally, reducing the compression residuals provides a direct means of controlling $\mathcal{D}_{\mathrm{gap}}^r$. Specifically, for fixed local updates and a given modality-wise rank allocation, the truncated SVD used by `Fed-MIAC` retains the leading singular components of each update, minimizing $\|e_{i,m}^r\|_F^2$ among all approximations of rank at most $\Lambda_{i,m}^r$. Since the gradient-based weights are nonnegative, this also minimizes the weighted residual sum in $\mathcal{D}_{\mathrm{gap}}^r$ for that allocation, without requiring knowledge of these weights. While truncated SVD itself minimizes reconstruction error for a given rank, the same amount of distortion across different modalities need not have equal effects on multimodal learning. To account for this asymmetry, `Fed-MIAC` incorporates modality importance when allocating the shared rank budget. The following section empirically validates the effectiveness of this importance-aware allocation.

6 EXPERIMENTS

We evaluate `Fed-MIAC` on four multimodal datasets covering vision–language, human activity sensing, clinical time series, and audio–text–video: (i) **CrisisMMD**(image & text), (ii) **KU-HAR**(accelerometer & gyroscope), (iii) **PTB-XL**(12-lead ECG), and (iv) **CMU-MOSEI**(audio & text & video)[3]. We run experiments with at least five seeds and report the mean performance.

6.1 EXPERIMENTAL SETTINGS

Baselines: Our evaluation includes `Fed-MIAC` (Ours), **SVDFed** (Wang et al., 2023), **MCB** (Li et al., 2024) and **MFedMC** (Yuan et al., 2026) by plug-in on top of two base FL training algorithms: **FedAvg** (McMahan et al., 2017) and the prototype-based multimodal FL method **MFCPL** (Le et al., 2025), which improves multimodal synergy by communicating class prototypes, and serves as a recent competitive and representative prototype-based MFL baseline. **SVDFed** is the unimodal approach to cut the transmission payload, while **MFedMC** and **MCB** are multimodal approaches to select one of the modality models to be transmitted for communication efficiency. Since **SVDFed** directly controls the communication ratio rather than the SVD rank, we report its single F1-score instead of providing rank-wise results for ranks 32, 16, 8, and 4. For a fair comparison, we apply the same client-sampling protocol as FedAvg for all methods.

FL settings: All experiments are conducted in a cross-device FL setup with a single central server and a dataset-dependent number of clients. For CrisisMMD, we simulate IID label distributions by partitioning the training split across 100 clients using a Dirichlet distribution with concentration parameter $\alpha = 5.0$. For KU-HAR, since each sample is associated with a subject ID, we treat each subject as a client and assign the subject's samples to its corresponding client. For PTB-XL, which contains multi-site hospital records, we treat each hospital as one client and allocate samples by site.

We split each dataset into training, validation, and test sets. FL is done on the training split, while validation/test splits are used solely for evaluation. We select the global checkpoint that achieves the best performance on validation and report its corresponding performance on the test set.

Communication budgets: To evaluate communication-efficient training, we control the allowed rank budgets $\Lambda \in \{32, 16, 8, 4\}$. Under a fixed Λ, we allocate the total rank budget across modalities according to modality importance, and perform truncated SVD for each modality using its assigned rank (smaller Λ implies more aggressive compression). In our results, $\Lambda = 32$ yields performance close to the uncompressed baseline, indicating that it provides sufficient capacity to preserve the original update information. To examine the effect of compression, we halve the rank to probe the accuracy–compression trade-off under progressively tighter budgets. For time-varying communication bandwidth, we also report a case of heterogeneous budgets across rounds in Appendix E.8.

We provide details of i) **SVDFed**, **MCB**, **MFedMC**, and **MFCPL**, ii) per-benchmark and per-modality model architectures, and iii) dataset splitting and hyperparameters in Appendix D.

6.2 PERFORMANCE EVALUATION

Performance gains over baselines: Table 1 reports the macro-F1 scores of all methods, with 95% confidence intervals provided in Appendix E.3. Across both base FL algorithms, i.e., FedAvg and MFCPL, `Fed-MIAC` consistently achieves better performance than all communication-efficient baselines over a wide range of communication budgets. This advantage remains evident under tight budgets of $\Lambda \in \{8, 4\}$, where the modality-selection baselines, i.e., MCB and MFedMC, exhibit substantial performance degradation relative to the corresponding uncompressed baseline. When compared with the second-best at $\Lambda = 8$, `Fed-MIAC` achieves gains up to $+4.42$ on CrisisMMD, $+5.51$ on KU-HAR, and $+3.17$ on PTB-XL in the FedAvg setting. Fed-MIAC also maintains its performance gains when combined with the prototype-based MFCPL algorithm, achieving improvements of $+5.02$, $+6.71$, and $+3.48$ macro-F1 points. Collectively, these results support fine-grained modality-wise compression as a more effective alternative to binary modality selection under constrained communication, with consistent gains across different base FL algorithms and budget levels.

[3]Our evaluation protocol and dataset preprocessing follow the FedMultimodal Benchmark GitHub (Feng et al., 2023). Due to the space constraints, we report additional experiments in Appendix E, including results on CMU-MOSEI (Appendix E.4), an extension to large models (Appendix E.5), and ablation studies.

Table 1: Test F1-score of base FL algorithms on CrisisMMD, KU-HAR, and PTB-XL.

Base FL	Dataset	CrisisMMD					KU-HAR					PTB-XL				
	Λ (Rank)	Full	32	16	8	4	Full	32	16	8	4	Full	32	16	8	4
FedAvg	SVDFed	50.30	37.92				66.74	52.59				61.83	57.13			
	MCB		46.49	46.67	43.72	37.63		59.13	56.28	56.35	49.76		58.10	57.93	57.03	55.65
	MFedMC		46.03	45.93	41.55	37.43		61.83	61.40	58.14	54.96		57.28	56.98	55.57	55.42
	Fed-MIAC		**49.54**	**49.33**	**48.14**	**41.07**		**67.01**	**65.66**	**63.65**	**58.21**		**61.97**	**61.54**	**60.20**	**59.21**
MFCPL	SVDFed	49.69	37.31				68.61	52.92				62.49	56.67			
	MCB		47.19	44.85	41.65	36.90		59.64	57.99	57.14	56.55		58.74	58.57	57.73	55.77
	MFedMC		44.68	43.83	39.48	36.31		60.26	59.88	58.63	55.67		57.81	58.10	56.76	55.25
	Fed-MIAC		**49.45**	**48.28**	**46.67**	**40.36**		**67.60**	**65.22**	**65.34**	**58.01**		**62.26**	**61.82**	**61.21**	**59.18**

Full: full rank / Λ: total rank budget across modalities

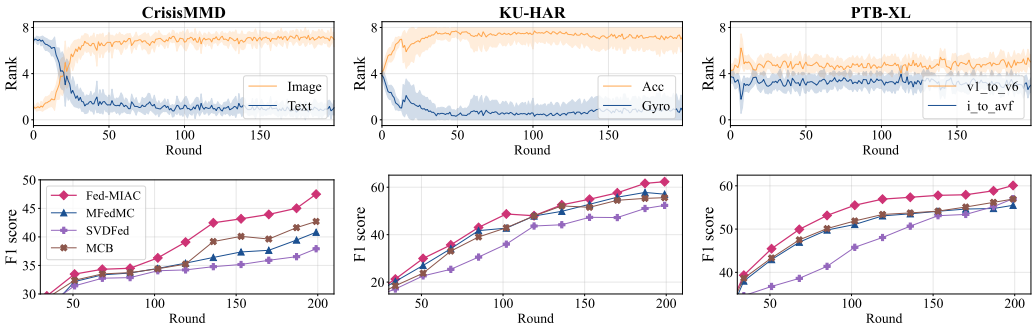

Figure 2: Behavior of rank allocation (top row) and FL convergence (bottom row) across datasets.

Rank allocation behavior: Beyond the aggregate performance gains, we further analyze how Fed-MIAC distributes a constrained rank budget across modalities, as this adaptive allocation constitutes the key mechanism underlying its performance advantage. The top row of Figure 2 shows the rank allocation behavior of Fed-MIAC under a fixed total rank budget of $\Lambda = 8$. The results show that the rank allocation patterns underlying the performance gains in Table 1 are not fixed, but vary substantially across datasets and evolve over communication rounds. Specifically, the rank allocation is moderately imbalanced toward the image modality on CrisisMMD, more strongly imbalanced toward the accelerometer modality on KU-HAR, and nearly balanced between the two ECG modality groups on PTB-XL. These patterns are driven jointly by the locally estimated relative modality importance and the singular-value structure of each modality-specific update. To isolate the former, we separately report the normalized modality-importance scores across rounds in Appendix E.6.

FL convergence: The bottom row of Figure 2 shows that Fed-MIAC consistently achieves favorable convergence behavior across all three datasets, while the gap over modality-selection baselines varies with the modality-importance profile. On KU-HAR, where the optimized rank allocation is strongly imbalanced toward one modality, the gap is relatively small because the effective allocation of Fed-MIAC becomes closer to binary modality selection. In contrast, on PTB-XL, where the two modalities receive nearly balanced ranks, Fed-MIAC exhibits a clearer advantage, as dropping either modality discards a substantial portion of complementary update information. Notably, in this balanced regime, the importance-agnostic SVDFed performs comparably to or better than the modality-selection baselines in later rounds, suggesting that hard modality dropping can be less effective when neither modality is clearly dominant. These results highlight the broader applicability of Fed-MIAC: unlike selection-based frameworks whose effectiveness depends on modality dominance, Fed-MIAC remains effective across diverse modality-importance profiles.

Communication–performance trade-off: Figure 3 compares the cumulative uplink communication cost required by FedAvg (an uncompressed baseline) and Fed-MIAC to achieve a given test performance, focusing on the uplink transmission that constitutes the target of our compression. To reach F1$=45$ on CrisisMMD and F1$=60$ on KU-HAR and PTB-XL, Fed-MIAC requires only 3.75%, 8.42%, and 7.65% of the corresponding FedAvg communication cost, respectively (8.45×10^8 vs. 2.25×10^{10}, 6.38×10^8 vs. 7.58×10^9, and 8.07×10^8 vs. 1.05×10^{10} bits). These results demonstrate substantial communication savings at matched performance across datasets.

Table 2: Wall-clock overhead of FedAvg and Fed-MIAC on Jetson TX2.

Dataset	Method	Comput. Ratios	Comput. Time (s)	Uplink Time (s)	Total Time (s)
CrisisMMD	FedAvg	–	118.52	3,754.23	3,872.75
	Fed-MIAC	[50.39%, 20.41%, 29.21%]	325.06	140.90	465.97 (**87.97%**↓)
KU-HAR	FedAvg	–	74.46	1,262.56	1,337.02
	Fed-MIAC	[51.48%, 18.97%, 29.54%]	236.87	106.29	343.16 (**74.33%**↓)
PTB-XL	FedAvg	–	404.73	1,757.70	2,162.43
	Fed-MIAC	[47.19%, 3.95%, 48.87%]	1,209.80	134.54	1,344.33 (**37.83%**↓)

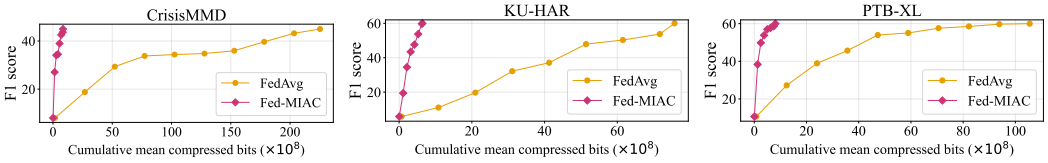

Figure 3: Communication efficiency on CrisisMMD (left), KU-HAR (middle), and PTB-XL (right)

Wall-clock overhead of SVD and SHAPE score on Jetson TX2: Although Fed-MIAC incurs additional computation for the SHAPE score and SVD, its practical efficiency should be assessed by total wall-clock time, including both computation and communication. We therefore execute the client-side pipeline on an NVIDIA Jetson TX2, whose specifications are provided in Appendix E.1. We measure the cumulative time to reach F1$=45$ on CrisisMMD, F1$=60$ on KU-HAR, and F1$=60$ on PTB-XL, reached at rounds 134/185, 112/184, and 136/196 for FedAvg/Fed-MIAC, respectively. Uplink time is estimated assuming a 6 Mbps transmission rate corresponding to IEEE 802.11g WiFi (IEEE, 2003), while the identical downlink cost is excluded. FedAvg computation includes local training, whereas Fed-MIAC additionally includes SHAPE-score computation and SVD.

As shown in Table 2, Fed-MIAC reduces the total time by 87.97%, 74.33%, and 37.83% relative to FedAvg on CrisisMMD, KU-HAR, and PTB-XL, respectively. These results show that the communication savings outweigh the additional SVD and SHAPE-score computation. In particular, PTB-XL incurs a substantial SHAPE-score cost (48.87% of the Fed-MIAC computation time), yet Fed-MIAC still achieves a lower end-to-end time because of its markedly reduced uplink cost.

Effectiveness of modality-importance-aware allocation: We additionally examine whether explicitly incorporating modality importance improves rank allocation beyond the singular-value information alone. Under the same total rank budget, we compare Fed-MIAC with two variants: Uniform, which evenly divides the rank budget across modalities, and w/o MIAC, which performs rank allocation using only the singular-value term while assigning equal importance to all modalities. To establish generality beyond SVD, we further verify the effectiveness of modality-aware resource allocation under a quantization-based compression scheme in Appendix E.2.

As shown in Table 3, incorporating modality importance achieves the best performance across all datasets and both rank budgets. In particular, Fed-MIAC consistently improves over w/o MIAC, showing that the singular-value structure alone is insufficient for determining how the shared communication budget should be distributed across modalities.

Table 3: Effectiveness of modality-importance-aware rank allocation across different budgets.

Dataset	CrisisMMD		KU-HAR		PTB-XL	
Methods	8	4	8	4	8	4
Uniform	44.63	40.38	61.85	56.49	58.67	58.03
+w/o MIAC	46.94	39.71	58.71	56.41	59.25	58.63
Fed-MIAC	**48.14**	**41.07**	**63.65**	**58.21**	**60.20**	**59.21**

7 Conclusion

In this work, we proposed Fed-MIAC, a modality-importance-aware compression framework tailored for multimodal federated learning. Fed-MIAC jointly considers two objectives—communication efficiency and preservation of multimodal synergy—and designs an adaptive compression strategy that allocates budgets according to modality importance. Experimental results demonstrate that Fed-MIAC outperforms existing methods across diverse datasets and budget settings, effectively maintaining multimodal performance under constrained resources.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on mathematical derivations and refine mathematical proofs. We did not use generative AI tools to develop theoretical models, formulate core mathematical claims, design experimental methodologies, implement methods, or interpret results; generating synthetic data sets, translation, or qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of new personal data. All experiments are conducted using publicly available benchmark datasets in accordance with their respective licenses and intended research use. The proposed methodology introduces no additional ethical concerns beyond those generally associated with machine learning and federated learning systems.

REPRODUCIBILITY STATEMENT

We provide the algorithmic details of `Fed-MIAC` in the main paper and Appendix A, and include additional implementation details in Appendix D, including hyperparameter settings, dataset descriptions, and preprocessing procedures. The assumptions and complete proofs for the theoretical results are also provided in Appendices B and C. Additional experimental results are included in Appendix E. We provide the source code necessary to reproduce the experiments as supplementary material.

REFERENCES

Firoj Alam, Ferda Ofli, and Muhammad Imran. CrisisMMD: Multimodal twitter datasets from natural disasters. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 2018. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14983.

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, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

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, 2019. doi: 10.1109/TPAMI.2018.2798607.

Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 4035–4045, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. URL https://doi.org/10.1145/3580305.3599825.

Pengbo Hu, Xingyu Li, and Yi Zhou. SHAPE: An unified approach to evaluate the contribution and cooperation of individual modalities. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3064–3070. International Joint Conferences on Artificial Intelligence Organization, 2022. doi: 10.24963/ijcai.2022/425.

IEEE. Ieee standard for information technology– local and metropolitan area networks– specific requirements– part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications: Further higher data rate extension in the 2.4 ghz band. *IEEE Std 802.11g-2003*

(Amendment to IEEE Std 802.11, 1999 Edn. (Reaff 2003) as amended by IEEE Stds 802.11a-1999, 802.11b-1999, 802.11b-1999/Cor 1-2001, and 802.11d-2001), pp. 1–104, 2003. doi: 10.1109/IEEESTD.2003.94282.

Huy Q. Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh N.H. Nguyen, Eui-Nam Huh, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, 122:103219, 2025. URL https://www.sciencedirect.com/science/article/pii/S1566253525002921.

Tan Li, Yanming Gong, Hai Liu, Zhen Chen, and Linqi Song. Communication-efficient multi-modal federated learning via dynamic client-modality matching. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 7899–7908, 2024. doi: 10.1109/BigData62323.2024.10825371.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Zhefeng Qiao, Xianghao Yu, Jun Zhang, and Khaled B Letaief. Communication-efficient federated learning with dual-side low-rank compression. *arXiv preprint arXiv:2104.12416*, 2021.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18250–18280. PMLR, 2022. URL https://proceedings.mlr.press/v162/qu22a.html.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2021–2031. PMLR, 2020. URL https://proceedings.mlr.press/v108/reisizadeh20a.html.

Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–318. Princeton University Press, Princeton, 1953.

Niloy Sikder and Abdullah-Al Nahid. KU-HAR: An open dataset for heterogeneous human activity recognition. *Pattern Recognition Letters*, 146:46–54, 2021. URL https://www.sciencedirect.com/science/article/pii/S0167865521000933.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15963–15974, June 2024.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, 2022. URL https://doi.org/10.13026/kfzx-aw45. Version 1.0.3.

Haolin Wang, Xuefeng Liu, Jianwei Niu, and Shaojie Tang. SVDFed: Enabling communication-efficient federated learning via singular-value-decomposition. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pp. 1–10, 2023. doi: 10.1109/INFOCOM53939.2023.10229042.

Yanmeng Wang, Yanqing Xu, Qingjiang Shi, and Tsung-Hui Chang. Quantized federated learning under transmission delay and outage constraints. *IEEE Journal on Selected Areas in Communications*, 40(1):323–341, 2022a. doi: 10.1109/JSAC.2021.3126081.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22802–22838. PMLR, 17–23 Jul 2022b.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.

Xu Yang, Zhuo Tang, Boyao Hao, Xiong Xiao, and Jiapeng Zhang. FedMPQ: Secure and efficient federated learning with multi-codebook product quantization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209457.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Hnk1WRMAYqg.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10790–10797, 2021. doi: 10.1609/aaai.v35i12.17289.

Liangqi Yuan, Dong-Jun Han, Su Wang, Devesh Upadhyay, and Christopher G Brinton. Communication-efficient multimodal federated learning: Joint modality and client selection. *IEEE Transactions on Mobile Computing*, 25(08):12780–12796, 2026. doi: 10.1109/TMC.2026.3673697. URL https://doi.ieeecomputersociety.org/10.1109/TMC.2026.3673697.

Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4):1476, 2022.

Song Zuo, Jinglei Wang, Xin Wang, Han Feng, Yize Zhao, Fadong Li, Shaobin Wei, Le Zhou, Zixu Zhao, Songjie Chen, et al. A dual-modal wearable ppg smartwatch with ai-enhanced correction for high-accuracy and continuous af burden assessment. *Advanced Science*, pp. e76627, 2026. doi: 10.1002/advs.76627.

A PSEUDO-CODE OF FED-MIAC

Algorithm 1 FL process of `Fed-MIAC`

Require: Initial global model w^0; local learning rate η; # local steps T; total rank budget Λ; temperature τ; # rounds R.
Ensure: Global model w^R

```
for r = 0 to R − 1 do
  Sample subset S^r ⊆ [N] of clients
  for all clients i ∈ S^r do
    w^r_{i,0} ← w^r
    for t = 0 to T − 1 do
      Sample mini-batch ξ_{i,t} ∼ D_i
      w^r_{i,t+1} ← w^r_{i,t} − η∇ℓ(w^r_{i,t}; ξ_{i,t})
    end for
    Δ^r_i ← w^r_{i,T} − w^r
    φ^r_i ← Calculate SHAPE score(w^r_{i,T}, D_i)
    α^r_i ← Softmax(φ^r_i, τ)
    for m = 1 to M do
      Compute SVD {σ_k(Δ^r_{i,m})}
      Compute importance-weighted singular values α^r_{i,m} · σ_k(Δ^r_{i,m})^2
    end for
    Select top-Λ of α^r_{i,m} · σ_k(Δ^r_{i,m})^2 across all (m, k)
    Set {Λ^r_{i,m}}_{m=1}^M ← # selected components belonging to modality m
    for m = 1, ..., M do
      (U^r_{i,m}, Σ^r_{i,m}, V^r_{i,m}) ← TSVD(Δ^r_{i,m}, Λ^r_{i,m})
    end for
    Upload {U^r_{i,m}, Σ^r_{i,m}, V^r_{i,m}}_{m=1}^M to the server
  end for
  Reconstruct Δ̂^r_i from {U^r_{i,m}, Σ^r_{i,m}, V^r_{i,m}}_{m=1}^M for each i ∈ S^r
  Δ̂^r ← Σ_{i∈S^r} (n_i / Σ_{k∈S^r} n_k) Δ̂^r_i
  w^{r+1} ← w^r + Δ̂^r
end for
```

We summarize the overall FL procedure (including the compression module) in Algorithm 1, presenting the end-to-end FL workflow from client sampling and local training to modality-aware compression and server aggregation.

B LAGRANGIAN SOLUTION OF IMPORTANCE-AWARE SVD RANK ALLOCATION

B.1 PROBLEM FORMULATION

Consider a matrix $W \in \mathbb{R}^{p\times q}$ to be compressed. Let $d := \min\{p, q\}$ and denote its singular values by

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0. \tag{6}$$

We compress W by truncated SVD with rank $\Lambda \in \{0, \ldots, d\}$. The resulting reconstruction is $W^{(\Lambda)}$, and the distortion is measured by the squared Frobenius residual:

$$D(W, \Lambda) := \|W - W^{(\Lambda)}\|_F^2 = \sum_{j>\Lambda} \sigma_j^2. \tag{7}$$

Now consider M modalities. For each modality $m \in \{1, \ldots, M\}$, let $W_m \in \mathbb{R}^{p_m\times q_m}$ be the update matrix and $d_m := \min\{p_m, q_m\}$. Given modality-importance weights $\alpha_m \geq 0$ and a total rank budget $\Lambda \in \mathbb{Z}_{\geq 0}$, we solve

$$\min_{\{\Lambda_m\in\{0,\ldots,d_m\}\}} \sum_{m=1}^{M} \alpha_m D(W_m, \Lambda_m) \quad \text{s.t.} \quad \sum_{m=1}^{M} \Lambda_m \leq \Lambda. \tag{8}$$

B.2 LAGRANGIAN FORMULATION

For $\lambda \geq 0$, define the Lagrangian

$$\mathcal{L}(\{\Lambda_m\}, \lambda) = \sum_{m=1}^{M} \alpha_m D(W_m, \Lambda_m) + \lambda\Big(\sum_{m=1}^{M} \Lambda_m - \Lambda\Big). \tag{9}$$

Fixing λ makes the problem separable across modalities:

$$\min_{\{\Lambda_m\in\{0,\ldots,d_m\}\}} \mathcal{L}(\{\Lambda_m\}, \lambda) = -\lambda\Lambda + \sum_{m=1}^{M} \min_{\Lambda_m\in\{0,\ldots,d_m\}} \Big(\alpha_m D(W_m, \Lambda_m) + \lambda\Lambda_m\Big). \tag{10}$$

B.3 SOLUTION FOR FIXED λ: A THRESHOLD RULE

For each modality m, define

$$\mathcal{L}_m(\Lambda_m) := \alpha_m D(W_m, \Lambda_m) + \lambda\Lambda_m, \qquad \Lambda_m \in \{0, \ldots, d_m\}. \tag{11}$$

Define the weighted marginal gain of the j-th singular component by

$$s_{m,j} := \alpha_m \sigma_{m,j}^2, \qquad j = 1, \ldots, d_m. \tag{12}$$

Since the singular values are non-increasing, $s_{m,1} \geq \cdots \geq s_{m,d_m} \geq 0$. Moreover,

$$\mathcal{L}_m(\Lambda_m) = \alpha_m \sum_{j=1}^{d_m} \sigma_{m,j}^2 + \sum_{j=1}^{\Lambda_m} (\lambda - s_{m,j}). \tag{13}$$

For $\Lambda_m \in \{0, \ldots, d_m - 1\}$, using Eq. (7), the one-step difference is

$$\begin{aligned} \mathcal{L}_m(\Lambda_m + 1) - \mathcal{L}_m(\Lambda_m) &= \alpha_m\big(D(W_m, \Lambda_m + 1) - D(W_m, \Lambda_m)\big) + \lambda \\ &= \alpha_m\Big(\sum_{j>\Lambda_m+1} \sigma_{m,j}^2 - \sum_{j>\Lambda_m} \sigma_{m,j}^2\Big) + \lambda \\ &= \lambda - s_{m,\Lambda_m+1}. \end{aligned} \tag{14}$$

We first consider a multiplier λ that does not coincide with any weighted marginal gain:

$$s_{m,j} \neq \lambda \qquad \text{for every } m \in [M] \text{ and } j \in \{1, \ldots, d_m\}. \tag{15}$$

Such a multiplier is constructed from the strict gap between two consecutive ordered gains in the next subsection. Then increasing the rank from Λ_m to $\Lambda_m + 1$ decreases $\mathcal{L}_m$ if and only if

$$s_{m,\Lambda_m+1} > \lambda, \tag{16}$$

and increases $\mathcal{L}_m$ if $s_{m,\Lambda_m+1} < \lambda$. Consequently, $\mathcal{L}_m$ strictly decreases while $s_{m,j} > \lambda$ and strictly increases after $s_{m,j} < \lambda$. Hence, the unique minimizer for fixed λ is

$$\Lambda_m^\star(\lambda) = \Big|\big\{j \in \{1, \dots, d_m\} : s_{m,j} > \lambda\big\}\Big|. \tag{17}$$

For an interior solution $1 \le \Lambda_m^\star(\lambda) < d_m$, this is equivalently characterized by

$$s_{m,\Lambda_m^\star(\lambda)} > \lambda > s_{m,\Lambda_m^\star(\lambda)+1}. \tag{18}$$

B.4 Global optimality: top-Λ selection view

For any feasible $\{\Lambda_m\}$, we rewrite the objective in Eq. (8) as

$$\begin{aligned}
\sum_{m=1}^{M} \alpha_m D(W_m, \Lambda_m) &= \sum_{m=1}^{M} \alpha_m \sum_{j>\Lambda_m} \sigma_{m,j}^2 \\
&= \sum_{m=1}^{M} \alpha_m \sum_{j=1}^{d_m} \sigma_{m,j}^2 - \sum_{m=1}^{M} \sum_{j=1}^{\Lambda_m} \alpha_m \sigma_{m,j}^2 \\
&= C - \sum_{m=1}^{M} \sum_{j=1}^{\Lambda_m} s_{m,j},
\end{aligned} \tag{19}$$

where $C := \sum_{m=1}^{M} \alpha_m \sum_{j=1}^{d_m} \sigma_{m,j}^2$ is a constant independent of $\{\Lambda_m\}$. Thus, minimizing Eq. (8) is equivalent to

$$\max_{\{\Lambda_m \in \{0,\dots,d_m\}\}} \sum_{m=1}^{M} \sum_{j=1}^{\Lambda_m} s_{m,j} \quad \text{s.t.} \quad \sum_{m=1}^{M} \Lambda_m \le \Lambda. \tag{20}$$

Denote all scores sorted in non-increasing order by

$$s_{(1)} \ge s_{(2)} \ge \cdots.$$

Assume that the Λ-th largest weighted marginal gain is strictly greater than the next largest gain:

$$s_{(\Lambda)} > s_{(\Lambda+1)}. \tag{21}$$

Choose any multiplier satisfying

$$s_{(\Lambda+1)} < \lambda < s_{(\Lambda)}. \tag{22}$$

Then exactly Λ scores are strictly larger than λ. Therefore, the modality-wise minimizer in Eq. (17) satisfy

$$\sum_{m=1}^{M} \Lambda_m^\star(\lambda) = |\{(m,j) : s_{m,j} > \lambda\}| = \Lambda. \tag{23}$$

The selected set $\{(m,j) : s_{m,j} > \lambda\}$ therefore consists precisely of the top-Λ weighted marginal gains. Since every feasible allocation selects at most Λ unit-cost components, its total selected gain cannot exceed $\sum_{k=1}^{\Lambda} s_{(k)}$. Hence, the threshold allocation maximizes Eq. (20).

This selection also corresponds to a valid truncated-SVD rank for every modality. Indeed, if component (m,j) is selected, then for every $\ell < j$, $s_{m,\ell} \ge s_{m,j} > \lambda$, so component (m,ℓ) is also selected. Thus, the selected components of modality m are exactly the first $\Lambda_m^\star(\lambda)$ singular components. Therefore, the top-Λ selection is feasible and globally optimal for Eq. (8).

Proposition B.1 (Optimal modality-importance-aware rank allocation)**.** *Let the singular values of W_m be $\{\sigma_k(W_m)\}_{k=1}^{d_m}$ in non-increasing order, where $d_m = \min\{p_m, q_m\}$, and define the weighted marginal gain as $s_{m,k} := \alpha_m \sigma_k(W_m)^2$. Let $s_{(1)} \ge s_{(2)} \ge \cdots$ be the ordered values of $\{s_{m,k}\}_{m,k}$. Assume that the Λ-th largest weighted marginal gain is strictly greater than the next largest gain, i.e., $s_{(\Lambda)} > s_{(\Lambda+1)}$. Then the unique optimal solution to Eq. (8) is obtained by selecting the Λ modality–rank pairs corresponding to $s_{(1)}, \dots, s_{(\Lambda)}$. The allocated rank for modality m is*

$$\Lambda_m^\star \;=\; \big|\{k :\ (m,k) \textit{ is selected}\}\big|.$$

Notation (Main Text and Appendix). In the Appendix, we denote an optimal allocation by $\{\Lambda_m^\star\}_{m=1}^M$. For simplicity, in the main text we drop the superscript and write Λ_m to denote the optimal solution whenever it is clear from the context.

C PROOFS OF THEORETICAL ANALYSIS (CONVERGENCE RATE)

C.1 SETTING

Let $w^r = (w_1^r, \ldots, w_M^r)$ be the global model at round r, where each block corresponds to one modality. The global objective is

$$F(w) = \sum_{i=1}^{N} p_i F_i(w), \qquad p_i \geq 0, \qquad \sum_{i=1}^{N} p_i = 1. \tag{24}$$

For client i, let $\Delta_i^r := -\eta \sum_{t=0}^{T-1} g_{i,t}^r$ denote the local model update, and let $\widehat{\Delta}_i^r$ be its compressed counterpart. Let $\Delta_{i,m}^r$ and $\widehat{\Delta}_{i,m}^r$ denote their respective modality-m blocks. Throughout this proof, inner products and norms involving update or compression-perturbation matrices are the Frobenius inner product and norm. We define the compression error by

$$\eta e_{i,m}^r \triangleq \widehat{\Delta}_{i,m}^r - \Delta_{i,m}^r, \qquad e_i^r = (e_{i,1}^r, \ldots, e_{i,M}^r), \qquad \bar{e}^r \triangleq \sum_{i=1}^{N} p_i e_i^r. \tag{25}$$

Equivalently,

$$\left\| \Delta_{i,m}^r - \widehat{\Delta}_{i,m}^r \right\|_F^2 = \eta^2 \left\| e_{i,m}^r \right\|_F^2.$$

As the learning rate decreases, the magnitude of local model updates typically decreases, and so does the corresponding parameter-space compression perturbation. We therefore define e_i^r as a learning-rate-normalized compression perturbation, such that $\widehat{\Delta}_i^r - \Delta_i^r = \eta e_i^r$.

The server update with client-side compression is

$$\begin{aligned} w^{r+1} &= w^r + \sum_{i=1}^{N} p_i \widehat{\Delta}_i^r \\ &= w^r + \sum_{i=1}^{N} p_i \Delta_i^r + \eta \sum_{i=1}^{N} p_i e_i^r \\ &= w^r - \eta \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r + \eta \sum_{i=1}^{N} p_i e_i^r \\ &= w^r - \eta \left(\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r - \bar{e}^r \right). \end{aligned} \tag{26}$$

C.2 ASSUMPTIONS

Assumption C.1 (L-smoothness and lower boundedness)**.** Each local objective $F_i(w)$ is L-Lipchitz continuous gradient, for all $i \in [N]$, i.e.,

$$\|\nabla F_i(v) - \nabla F_i(u)\| \leq L\|v - u\|, \quad \forall u, v.$$

Moreover, there exists $F^\star > -\infty$ such that $F_i(w) \geq F^\star$ for every w.

Assumption C.2 (Bounds of the stochastic gradients)**.** The stochastic gradient $\nabla F_i(w, \zeta_i)$ with batch ζ_i is an unbiased estimator of $\nabla F_i(w)$ with variance bounded by σ_l^2 for all $i \in [N]$, i.e.,

$$\mathbb{E}_{\zeta_i} \|\nabla F_i(w, \zeta_i) - \nabla F_i(w)\|^2 \leq \sigma_l^2.$$

Assumption C.3 (Bounds of gradients divergence)**.** The global variability of the local gradient is bounded by σ_g^2 for all $i \in [N]$, i.e.,

$$\|\nabla F_i(w^r) - \nabla F(w^r)\|^2 \leq \sigma_g^2.$$

Assumption C.4 (Bounds of squared gradient norm)**.** The expected squared norm of gradients is uniformly bounded for all $i \in [N]$, $t \in [T]$ and $r \in [R]$, i.e.,

$$\mathbb{E} \left\| \nabla F_i(w_{i,t}^r, \zeta_{i,t}^r) \right\|^2 \leq \mathcal{G}^2.$$

Assumption C.5 (Bounded compression error)**.** For all clients $i \in [N]$, modalities $m \in [M]$, and communication rounds r, the compression error is bounded as

$$\|e_{i,m}^r\|^2 \le e^2,$$

where $e > 0$ is a finite constant.

C.3 One-step descent

From the server update,

$$w^{r+1} = w^r - \eta \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r + \eta \sum_{i=1}^{N} p_i e_i^r. \tag{27}$$

Using Assumption 1 with

$$x = w^r, \qquad y = w^{r+1},$$

we have

$$F(w^{r+1}) \le F(w^r) + \left\langle \nabla F(w^r), w^{r+1} - w^r \right\rangle + \frac{L}{2} \left\| w^{r+1} - w^r \right\|^2. \tag{28}$$

Since

$$w^{r+1} - w^r = -\eta \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r + \eta \sum_{i=1}^{N} p_i e_i^r, \tag{29}$$

we obtain

$$\begin{aligned} F(w^{r+1}) \le F(w^r) &+ \left\langle \nabla F(w^r), -\eta \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r + \eta \sum_{i=1}^{N} p_i e_i^r \right\rangle \\ &+ \frac{L}{2} \left\| -\eta \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r + \eta \sum_{i=1}^{N} p_i e_i^r \right\|^2. \end{aligned} \tag{30}$$

Expanding the inner-product term,

$$\begin{aligned} F(w^{r+1}) \le F(w^r) &- \eta \left\langle \nabla F(w^r), \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\rangle \\ &+ \eta \left\langle \nabla F(w^r), \sum_{i=1}^{N} p_i e_i^r \right\rangle \\ &+ \frac{L}{2} \left\| \eta \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r - \eta \sum_{i=1}^{N} p_i e_i^r \right\|^2. \end{aligned} \tag{31}$$

$$\begin{aligned} F(w^{r+1}) \le &\underbrace{F(w^r) - \eta \left\langle \nabla F(w^r), \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\rangle + \frac{L\eta^2}{2} \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2}_{\text{terms without compression error}} \\ &+ \underbrace{\left[\eta \left\langle \nabla F(w^r), \sum_{i=1}^{N} p_i e_i^r \right\rangle - L\eta^2 \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \sum_{i=1}^{N} p_i e_i^r \right\rangle + \frac{L\eta^2}{2} \left\| \sum_{i=1}^{N} p_i e_i^r \right\|^2 \right]}_{\text{compression-error-dependent terms}}. \end{aligned} \tag{32}$$

C.4 Terms without compression error

We now bound the **terms without compression error** in Eq. (32) directly, without introducing auxiliary symbols for the aggregated local direction or its deviation. The polarization identity gives

$$
\begin{aligned}
&F(w^r) - \eta \left\langle \nabla F(w^r), \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\rangle + \frac{L\eta^2}{2} \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2 \\
&= F(w^r) - \frac{\eta T}{2} \left\| \nabla F(w^r) \right\|^2 - \frac{\eta}{2T}(1 - L\eta T) \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2 + \frac{\eta}{2T} \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r - T\nabla F(w^r) \right\|^2 \\
&= F(w^r) - \frac{\eta T}{2} \left\| \nabla F(w^r) \right\|^2 - \frac{\eta}{2T}(1 - L\eta T) \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2 \\
&\quad + \frac{\eta}{2T} \left(\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \|^2 - 2 \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, T\nabla F(w^r) \right\rangle + \| T\nabla F(w^r) \|^2 \right) \\
&= F(w^r) - \eta \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \nabla F(w^r) \right\rangle + \frac{L\eta^2}{2} \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2
\end{aligned} \tag{33}
$$

C.4.1 Upper Bound on the Inner Product Between Aggregated Local and Global Gradients

For each client and local step, add and subtract the two deterministic gradients $\nabla F_i(w_{i,t}^r)$ and $\nabla F_i(w^r)$:

$$
g_{i,t}^r = \left[g_{i,t}^r - \nabla F_i(w_{i,t}^r) \right] + \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r) \right] + \nabla F_i(w^r). \tag{34}
$$

Substituting Eq. (34) into the inner product of Eq. (33) gives

$$
\begin{aligned}
&- \eta \mathbb{E} \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \nabla F(w^r) \right\rangle \\
&\quad = -\eta \mathbb{E} \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[g_{i,t}^r - \nabla F_i(w_{i,t}^r) \right], \nabla F(w^r) \right\rangle \\
&\qquad - \eta \mathbb{E} \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r) \right], \nabla F(w^r) \right\rangle \\
&\qquad - \eta \mathbb{E} \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \nabla F_i(w^r), \nabla F(w^r) \right\rangle.
\end{aligned} \tag{35}
$$

The first inner product on the right-hand side of Eq. (35) vanishes. Indeed, Assumption 2 implies that, conditional on $w_{i,t}^r$,

$$
\mathbb{E}\left[g_{i,t}^r - \nabla F_i(w_{i,t}^r) \mid w_{i,t}^r \right] = 0,
$$

whereas $\nabla F(w^r)$ is fixed before the mini-batches in round r are sampled. For the third inner product, full participation gives

$$
\begin{aligned}
\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \nabla F_i(w^r) &= T \sum_{i=1}^{N} p_i \nabla F_i(w^r) \\
&= T\nabla F(w^r).
\end{aligned} \tag{36}
$$

Consequently,

$$
\begin{aligned}
&-\eta\mathbb{E}\left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \nabla F(w^r) \right\rangle \\
&= -\eta T\mathbb{E}\left\|\nabla F(w^r)\right\|^2 - \eta\mathbb{E}\left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right], \nabla F(w^r) \right\rangle \\
&\le -\eta T\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + \eta\mathbb{E}\left[\left\|\nabla F(w^r)\right\| \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right] \right\|\right] \\
&\le -\frac{3\eta T}{4}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + \frac{\eta}{T}\mathbb{E}\left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right] \right\|^2 . \qquad (37)
\end{aligned}
$$

The final inequality uses

$$ab \le \frac{T}{4}a^2 + \frac{1}{T}b^2$$

with $a = \|\nabla F(w^r)\|$ and

$$b = \left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right] \right\|.$$

C.4.2 Upper Bound on the Local Gradient Deviation from the Global Model

By applying Jensen's inequality and L-smoothness to $\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right]\right\|^2$ in Eq. (37),

$$
\begin{aligned}
\mathbb{E}\left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right] \right\|^2 &= T^2\mathbb{E}\left\| \sum_{i=1}^{N} \sum_{t=0}^{T-1} \frac{p_i}{T} \left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right] \right\|^2 && (38) \\
&\le T \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right\|^2 && (39) \\
&\le L^2 T \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\|w_{i,t}^r - w^r\|^2. && (40)
\end{aligned}
$$

The local recursion and Cauchy–Schwarz give, for every t,

$$w_{i,t}^r - w^r = -\eta \sum_{s=0}^{t-1} g_{i,s}^r, \qquad \|w_{i,t}^r - w^r\|^2 \le \eta^2 t \sum_{s=0}^{t-1} \|g_{i,s}^r\|^2.$$

Hence,

$$
\begin{aligned}
\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\|w_{i,t}^r - w^r\|^2 &\le \eta^2 \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} t \sum_{s=0}^{t-1} \mathbb{E}\|g_{i,s}^r\|^2 && (41) \\
&\le \eta^2 T^2 \sum_{i=1}^{N} p_i \sum_{s=0}^{T-1} \mathbb{E}\|g_{i,s}^r\|^2. && (42)
\end{aligned}
$$

By conditional unbiasedness in Assumption 2, the stochastic-gradient noise has zero cross term with $\nabla F_i(w_{i,s}^r)$. Therefore,

$$
\begin{aligned}
\mathbb{E}\|g_{i,s}^r\|^2 &= \mathbb{E}\left\|g_{i,s}^r - \nabla F_i(w_{i,s}^r) + \nabla F_i(w_{i,s}^r)\right\|^2 && (43) \\
&= \mathbb{E}\left\|g_{i,s}^r - \nabla F_i(w_{i,s}^r)\right\|^2 + \mathbb{E}\|\nabla F_i(w_{i,s}^r)\|^2 && (44) \\
&\le \sigma_l^2 + 2\mathbb{E}\|\nabla F_i(w_{i,s}^r) - \nabla F_i(w^r)\|^2 + 2\mathbb{E}\|\nabla F_i(w^r)\|^2 && (45) \\
&\le \sigma_l^2 + 2L^2\mathbb{E}\|w_{i,s}^r - w^r\|^2 + 2\mathbb{E}\|\nabla F_i(w^r)\|^2. && (46)
\end{aligned}
$$

Moreover, the variance decomposition and Assumption 3 imply

$$\sum_{i=1}^{N} p_i \|\nabla F_i(w^r)\|^2 = \|\nabla F(w^r)\|^2 + \sum_{i=1}^{N} p_i \|\nabla F_i(w^r) - \nabla F(w^r)\|^2 \tag{47}$$

$$\leq \|\nabla F(w^r)\|^2 + \sigma_g^2. \tag{48}$$

Combining the last two displays and summing over (i, s) yields

$$\sum_{i=1}^{N} p_i \sum_{s=0}^{T-1} \mathbb{E}\|g_{i,s}^r\|^2 \leq 2L^2 \sum_{i=1}^{N} p_i \sum_{s=0}^{T-1} \mathbb{E}\|w_{i,s}^r - w^r\|^2 + T\left(2\mathbb{E}\|\nabla F(w^r)\|^2 + 2\sigma_g^2 + \sigma_l^2\right).$$

Substitution into Eq. (42) gives

$$\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\|w_{i,t}^r - w^r\|^2 \leq 2L^2\eta^2T^2 \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\|w_{i,t}^r - w^r\|^2 \tag{49}$$

$$+ \eta^2T^3\left(2\mathbb{E}\|\nabla F(w^r)\|^2 + 2\sigma_g^2 + \sigma_l^2\right), \tag{50}$$

$$\left(1 - 2L^2\eta^2T^2\right) \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\|w_{i,t}^r - w^r\|^2 \leq \eta^2T^3\left(2\mathbb{E}\|\nabla F(w^r)\|^2 + 2\sigma_g^2 + \sigma_l^2\right). \tag{51}$$

Under $L\eta T \leq 1/8$, $2L^2\eta^2T^2 \leq 1/32$, whence $(1 - 2L^2\eta^2T^2)^{-1} \leq 2$. We conclude that

$$\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} \mathbb{E}\|w_{i,t}^r - w^r\|^2 \leq \frac{\eta^2T^3}{1 - 2L^2\eta^2T^2}\left(2\mathbb{E}\|\nabla F(w^r)\|^2 + 2\sigma_g^2 + \sigma_l^2\right) \tag{52}$$

$$\leq 2\eta^2T^3\left(2\mathbb{E}\|\nabla F(w^r)\|^2 + 2\sigma_g^2 + \sigma_l^2\right) \tag{53}$$

$$\mathbb{E}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1}\left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right]\right\|^2 \leq 2L^2\eta^2T^4\left(2\mathbb{E}\|\nabla F(w^r)\|^2 + 2\sigma_g^2 + \sigma_l^2\right). \tag{54}$$

C.4.3 Upper bound of aggregated local gradient

For $\frac{L\eta^2}{2}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r\right\|^2$ in Eq. (33), we first decompose the aggregated local direction into its global-gradient, stochastic-noise, and local-update-drift components:

$$\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r = T\nabla F(w^r) + \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1}\left[g_{i,t}^r - \nabla F_i(w_{i,t}^r)\right]$$

$$+ \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1}\left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right]. \tag{55}$$

This follows by adding and subtracting $\nabla F_i(w_{i,t}^r)$ and $\nabla F_i(w^r)$, then using $\sum_i p_i \nabla F_i(w^r) = \nabla F(w^r)$.

Stochastic-noise component. Since the weights p_i/T sum to one over (i, t), Jensen's inequality and Assumption 2 give

$$\mathbb{E}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1}\left[g_{i,t}^r - \nabla F_i(w_{i,t}^r)\right]\right\|^2$$

$$= T^2\mathbb{E}\left\|\sum_{i=1}^{N}\sum_{t=0}^{T-1} \frac{p_i}{T}\left[g_{i,t}^r - \nabla F_i(w_{i,t}^r)\right]\right\|^2$$

$$\leq T^2 \sum_{i=1}^{N}\sum_{t=0}^{T-1} \frac{p_i}{T}\mathbb{E}\left\|g_{i,t}^r - \nabla F_i(w_{i,t}^r)\right\|^2$$

$$\leq T^2\sigma_l^2. \tag{56}$$

Combining the three components. Applying $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ first to the stochastic-noise component and the sum of the other two components, and then to the global-gradient and local-update-drift components, gives

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r\right\|^2 &\le 4T^2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2\mathbb{E}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1}\left[g_{i,t}^r - \nabla F_i(w_{i,t}^r)\right]\right\|^2 \\ &\quad + 4\mathbb{E}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1}\left[\nabla F_i(w_{i,t}^r) - \nabla F_i(w^r)\right]\right\|^2 \\ &\le 4T^2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2T^2\sigma_l^2 \\ &\quad + 8L^2\eta^2T^4\left(2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2\sigma_g^2 + \sigma_l^2\right). \end{aligned} \tag{57}$$

The last inequality uses Eq. (56) and the D.4.2 result Eq. (54). Multiplying Eq. (57) by $L\eta^2/2$ gives the desired upper bound:

$$\begin{aligned}&\frac{L\eta^2}{2}\mathbb{E}\left\|\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r\right\|^2 \\ &\le \left(2L\eta^2T^2 + 8L^3\eta^4T^4\right)\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \\ &\quad + \left(L\eta^2T^2 + 4L^3\eta^4T^4\right)\sigma_l^2 + 8L^3\eta^4T^4\sigma_g^2. \end{aligned} \tag{58}$$

C.4.4 Convergence in no compression

We now combine the preceding bounds. First, substituting Eq. (54) into the final term of Eq. (37) gives

$$\begin{aligned}&-\eta\mathbb{E}\left\langle\sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \nabla F(w^r)\right\rangle \\ &\le -\left(\frac{3\eta T}{4} - 4L^2\eta^3T^3\right)\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \\ &\quad + 2L^2\eta^3T^3\sigma_l^2 + 4L^2\eta^3T^3\sigma_g^2. \end{aligned} \tag{59}$$

Next, apply Eq. (58) to the second-order term in Eq. (33). By Eq. (33) and Eq. (59), we obtain

$$\begin{aligned}\mathbb{E}[F(w^{r+1})] \le \mathbb{E}[F(w^r)] &- \left(\frac{3\eta T}{4} - 2L\eta^2T^2 - 4L^2\eta^3T^3 - 8L^3\eta^4T^4\right)\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \\ &+ \left(L\eta^2T^2 + 2L^2\eta^3T^3 + 4L^3\eta^4T^4\right)\sigma_l^2 \\ &+ \left(4L^2\eta^3T^3 + 8L^3\eta^4T^4\right)\sigma_g^2. \end{aligned} \tag{60}$$

Under $L\eta T \le 1/8$,

$$2L\eta T + 4(L\eta T)^2 + 8(L\eta T)^3 \le \frac{1}{4} + \frac{1}{16} + \frac{1}{64} = \frac{21}{64}.$$

It follows that

$$\frac{3\eta T}{4} - 2L\eta^2T^2 - 4L^2\eta^3T^3 - 8L^3\eta^4T^4 \ge \left(\frac{3}{4} - \frac{21}{64}\right)\eta T = \frac{27\eta T}{64} \ge \frac{\eta T}{4}.$$

Thus,

$$\begin{aligned}\mathbb{E}[F(w^{r+1})] \le \mathbb{E}[F(w^r)] &- \frac{\eta T}{4}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \\ &+ \left(L\eta^2T^2 + 2L^2\eta^3T^3 + 4L^3\eta^4T^4\right)\sigma_l^2 \\ &+ \left(4L^2\eta^3T^3 + 8L^3\eta^4T^4\right)\sigma_g^2. \end{aligned} \tag{61}$$

C.5 TERMS WITH COMPRESSION ERROR

We analyze all three compression-error-dependent terms in Eq. (32) using the revised first- and second-term bounds.

First term in Eq. (32) For $\sum_{m=1}^{M} \|\nabla_{w_m} F(w^r)\|_F > 0$, the quantities

$$p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F}$$

form a collection of weights that sums to one over (i, m). Jensen's inequality therefore gives

$$\begin{aligned}
&\left(\sum_{i=1}^{N} \sum_{m=1}^{M} p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F} \left\|e_{i,m}^r\right\|_F \right)^2 \\
&\quad \le \sum_{i=1}^{N} \sum_{m=1}^{M} p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F} \left\|e_{i,m}^r\right\|_F^2 . \qquad (62)
\end{aligned}$$

Blockwise Cauchy–Schwarz and Eq. (62) give

$$\begin{aligned}
&\eta \left\langle \nabla F(w^r), \sum_{i=1}^{N} p_i e_i^r \right\rangle \\
&\le \eta \left(\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F \right) \times \sum_{i=1}^{N} \sum_{m=1}^{M} p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F} \left\|e_{i,m}^r\right\|_F \\
&\le \eta \left(\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F \right) \times \left(\sum_{i=1}^{N} \sum_{m=1}^{M} p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{\sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F} \left\|e_{i,m}^r\right\|_F^2 \right)^{1/2} \\
&= \eta \left[G^2 \sum_{i=1}^{N} \sum_{m=1}^{M} p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{G} \left\|e_{i,m}^r\right\|_F^2 \right]^{1/2} , \qquad (63)
\end{aligned}$$

where $G := \sum_{k=1}^{M} \|\nabla_{w_k} F(w^r)\|_F$. When $\sum_{m=1}^{M} \|\nabla_{w_m} F(w^r)\|_F = 0$, the left-hand side is zero. In this case, we define the right-hand side of Eq. (63) to be zero, so the inequality remains well defined.

Second term in Eq. (32) By Cauchy–Schwarz in the joint randomness of stochastic gradients and compression errors,

$$\mathbb{E}\left[-L\eta^2 \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \bar{e}^r \right\rangle \right] \le L\eta^2 \left(\mathbb{E}\left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2 \right)^{1/2} \left(\mathbb{E}\left\|\bar{e}^r\right\|^2 \right)^{1/2} \qquad (64)$$

To bound the first factor, we use Eq. (57),

$$\begin{aligned}
\mathbb{E}\left\| \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r \right\|^2 &\le 4T^2 \mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2T^2\sigma_l^2 \\
&\quad + 8L^2\eta^2T^4 \left(2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2\sigma_g^2 + \sigma_l^2 \right) . \qquad (65)
\end{aligned}$$

Combining Eq. (64) and Eq. (65) yields

$$\begin{aligned}
\mathbb{E}\left[-L\eta^2 \left\langle \sum_{i=1}^{N} p_i \sum_{t=0}^{T-1} g_{i,t}^r, \bar{e}^r \right\rangle \right] &\le L\eta^2 \Bigg[4T^2 \mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2T^2\sigma_l^2 \\
&\quad + 8L^2\eta^2T^4 \left(2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2\sigma_g^2 + \sigma_l^2 \right) \Bigg]^{1/2} \left(\mathbb{E}\left\|\bar{e}^r\right\|^2 \right)^{1/2} . \qquad (66)
\end{aligned}$$

Summary of the terms in Eq. (32) Let $\mathcal{R}^r_{\text{comp}}$ denote the entire second underbraced quantity in Eq. (32), and define

$$\mathcal{D}^r_{\text{gap}} \triangleq \mathbb{E}\left[G^2 \times \sum_{i=1}^{N}\sum_{m=1}^{M} p_i \frac{\|\nabla_{w_m} F(w^r)\|_F}{G} \left\|e^r_{i,m}\right\|_F^2\right],$$

$$\mathcal{M}^r_{\text{dir}} \triangleq \left[\left(4T^2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2T^2\sigma_l^2 + 8L^2\eta^2T^4\left(2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2\sigma_g^2 + \sigma_l^2\right)\right)\mathbb{E}\left\|\bar{e}^r\right\|^2\right]^{1/2}$$

$$\mathcal{E}^r_{\text{agg}} \triangleq \mathbb{E}\left\|\bar{e}^r\right\|^2. \tag{67}$$

The integrand defining $\mathcal{D}^r_{\text{gap}}$ is set to zero when $G = 0$. Taking expectations in Eq. (63) and applying Cauchy–Schwarz, then combining with Eq. (66) and the remaining quadratic term, gives

$$\mathcal{E}^r_{\text{comp}} \triangleq \mathbb{E}\left[\mathcal{R}^r_{\text{comp}}\right] \le \eta\left(\mathcal{D}^r_{\text{gap}}\right)^{1/2} + L\eta^2\mathcal{M}^r_{\text{dir}} + \frac{L\eta^2}{2}\mathcal{E}^r_{\text{agg}}. \tag{68}$$

The three terms on the right-hand side arise, respectively, from the importance-weighted gradient–residual term in Eq. (63), the local-direction–aggregate-residual cross term in Eq. (66), and the remaining aggregate-residual quadratic term in Eq. (32). Every term on the right-hand side vanishes under exact communication, i.e., when $e^r_i = 0$ for all clients.

C.6 Multi-round convergence analysis

Combining the no-compression bound in Eq. (61) with the compression-error bound Eq. (68), we obtain, for $L\eta T \le 1/8$,

$$\begin{aligned}\mathbb{E}[F(w^{r+1})] \le \mathbb{E}[F(w^r)] &- \frac{\eta T}{4}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \\ &+ \left[L\eta^2T^2 + 2L^2\eta^3T^3(1+2L\eta T)\right]\sigma_l^2 \\ &+ 4L^2\eta^3T^3(1+2L\eta T)\sigma_g^2 + \mathcal{E}^r_{\text{comp}},\end{aligned} \tag{69}$$

where, by Eq. (68),

$$\mathcal{E}^r_{\text{comp}} \le \eta\left(\mathcal{D}^r_{\text{gap}}\right)^{1/2} + L\eta^2\mathcal{M}^r_{\text{dir}} + \frac{L\eta^2}{2}\mathcal{E}^r_{\text{agg}}. \tag{70}$$

Rearranging Eq. (69), summing over $r = 0, \ldots, R-1$, using $F(w^R) \ge F^\star$, and dividing by $\eta TR/4$ gives

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \le \frac{4\left(F(w^0) - F^\star\right)}{\eta TR} &+ \left[4L\eta T + 8L^2\eta^2T^2(1+2L\eta T)\right]\sigma_l^2 \\ &+ 16L^2\eta^2T^2(1+2L\eta T)\sigma_g^2 \\ &+ \frac{4}{\eta TR}\sum_{r=0}^{R-1}\mathcal{E}^r_{\text{comp}}.\end{aligned} \tag{71}$$

Substituting Eq. (70) yields

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \le \frac{4\left(F(w^0) - F^\star\right)}{\eta TR} &+ \left[4L\eta T + 8L^2\eta^2T^2(1+2L\eta T)\right]\sigma_l^2 \\ &+ 16L^2\eta^2T^2(1+2L\eta T)\sigma_g^2 \\ &+ \frac{4}{TR}\sum_{r=0}^{R-1}\left(\mathcal{D}^r_{\text{gap}}\right)^{1/2} + \frac{4L\eta}{TR}\sum_{r=0}^{R-1}\mathcal{M}^r_{\text{dir}} + \frac{2L\eta}{TR}\sum_{r=0}^{R-1}\mathcal{E}^r_{\text{agg}}.\end{aligned} \tag{72}$$

Following the standard diminishing-step-size convention in prior FedAvg convergence analyses (Qu et al., 2022), setting $\eta = 1/\sqrt{R}$ makes $L\eta T \le 1/8$ hold whenever $R \ge 64L^2T^2$. In this case,

Eq. (72) becomes

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla F(w^r)\right\|^2 \le \frac{4\left(F(w^0)-F^\star\right)}{T\sqrt{R}} + \left[\frac{4LT}{\sqrt{R}} + \frac{8L^2T^2}{R}\left(1+\frac{2LT}{\sqrt{R}}\right)\right]\sigma_l^2 + \frac{16L^2T^2}{R}\left(1+\frac{2LT}{\sqrt{R}}\right)\sigma_g^2 + \frac{4}{TR}\sum_{r=0}^{R-1}\left(\mathcal{D}_{\text{gap}}^r\right)^{1/2} + \frac{L}{TR^{3/2}}\sum_{r=0}^{R-1}\left(4\mathcal{M}_{\text{dir}}^r + 2\mathcal{E}_{\text{agg}}^r\right). \tag{73}$$

Every compression-dependent term vanishes under exact communication, i.e., when $e_i^r = 0$ for all i and r. Without compression, we define the convergence term as

$$\mathcal{B}_{\text{FL}} \triangleq \frac{4\left(F(w^0)-F^\star\right)}{T\sqrt{R}} + \left[\frac{4LT}{\sqrt{R}} + \frac{8L^2T^2}{R}\left(1+\frac{2LT}{\sqrt{R}}\right)\right]\sigma_l^2 + \frac{16L^2T^2}{R}\left(1+\frac{2LT}{\sqrt{R}}\right)\sigma_g^2$$

For fixed $L, T, \sigma_l^2, \sigma_g^2$, the baseline term satisfies $\mathcal{B}_{\text{FL}} = O(R^{-1/2})$.

C.7 Origins of the Compression Error Terms

Below is a compact map from each compression quantity to the term from which it is derived:

- $\mathcal{D}_{\text{gap}}^r$ is derived from the gradient–compression-residual term $\eta\langle\nabla F(w^r), \bar{e}^r\rangle$ in Eq. (32). Its importance-weighted bound is established in Eq. (63); after taking expectations, it contributes $\eta(\mathcal{D}_{\text{gap}}^r)^{1/2}$ to Eq. (68) and $\frac{4}{TR}(\mathcal{D}_{\text{gap}}^r)^{1/2}$ to the final bound.

$$\mathcal{D}_{\text{gap}}^r \triangleq \mathbb{E}\left[G^2 \times \sum_{i=1}^{N}\sum_{m=1}^{M} p_i \frac{\left\|\nabla_{w_m} F(w^r)\right\|_F}{G}\left\|e_{i,m}^r\right\|_F^2\right]. \tag{74}$$

  As above, the integrand is defined to be zero when $G = 0$.

- $\mathcal{M}_{\text{dir}}^r$ is derived from the cross term $-L\eta^2\langle\sum_i p_i \sum_{t=0}^{T-1} g_{i,t}^r, \bar{e}^r\rangle$ in Eq. (32). Combining the aggregated-direction second-moment bound Eq. (65) with Cauchy–Schwarz gives Eq. (66); hence, it contributes $L\eta^2\mathcal{M}_{\text{dir}}^r$ to Eq. (68) and $\frac{4L\eta}{TR}\mathcal{M}_{\text{dir}}^r$ to the final bound.

$$\mathcal{M}_{\text{dir}}^r \triangleq \Big[4T^2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2T^2\sigma_l^2 + 8L^2\eta^2T^4\left(2\mathbb{E}\left\|\nabla F(w^r)\right\|^2 + 2\sigma_g^2 + \sigma_l^2\right)\Big]^{1/2}\left(\mathbb{E}\left\|\bar{e}^r\right\|^2\right)^{1/2} \tag{75}$$

- $\mathcal{E}_{\text{agg}}^r$ is derived directly from the quadratic aggregate-residual term $\frac{L\eta^2}{2}\|\bar{e}^r\|^2$ in Eq. (32). Taking expectations yields $\frac{L\eta^2}{2}\mathcal{E}_{\text{agg}}^r$ in Eq. (68) and $\frac{2L\eta}{TR}\mathcal{E}_{\text{agg}}^r$ in the final bound.

$$\mathcal{E}_{\text{agg}}^r \triangleq \mathbb{E}\left\|\bar{e}^r\right\|^2 \tag{76}$$

C.8 Upper bound of the Compression Error Terms

We next show that the two compression quantities multiplied by $R^{-3/2}$ in Eq. (73) are uniformly bounded. Hence, although the quantities themselves need not decrease with the number of rounds, their contributions to the final convergence bound decay as $O(R^{-1/2})$.

Proposition C.6 (Uniform bounds for the aggregated compression terms)**.** *Under Assumptions C.1–C.4 and Assumption C.5, for every communication round r,*

$$\mathcal{E}_{\text{agg}}^r \le Me^2, \tag{77}$$

$$\mathcal{M}_{\text{dir}}^r \le e\sqrt{M}\left[4T^2\mathcal{G}^2 + 2T^2\sigma_l^2 + 8L^2\eta^2T^4\left(2\mathcal{G}^2 + 2\sigma_g^2 + \sigma_l^2\right)\right]^{1/2}. \tag{78}$$

Moreover, if $L\eta T \le 1/8$, then

$$\mathcal{M}^r_{\text{dir}} \le Te\sqrt{M}\left(\frac{17}{4}\mathcal{G}^2 + \frac{1}{4}\sigma_g^2 + \frac{17}{8}\sigma_l^2\right)^{1/2} \triangleq \overline{\mathcal{M}}_{\text{dir}}. \tag{79}$$

Consequently, the contribution of these two terms in Eq. (73) satisfies

$$\frac{L}{TR^{3/2}}\sum_{r=0}^{R-1}\left(4\mathcal{M}^r_{\text{dir}} + 2\mathcal{E}^r_{\text{agg}}\right) \le \frac{L}{T\sqrt{R}}\left(4\overline{\mathcal{M}}_{\text{dir}} + 2Me^2\right) = O(R^{-1/2}). \tag{80}$$

Here, $L, T, M, e, \mathcal{G}, \sigma_l$, and σ_g are treated as constants independent of R.

Proof. Because the modality blocks occupy disjoint parameter coordinates, Assumption C.5 gives

$$\|e_i^r\|^2 = \sum_{m=1}^{M}\|e_{i,m}^r\|_F^2 \le Me^2. \tag{81}$$

Since $p_i \ge 0$ and $\sum_i p_i = 1$, convexity of the squared norm yields

$$\mathcal{E}^r_{\text{agg}} = \mathbb{E}\left\|\sum_{i=1}^{N} p_i e_i^r\right\|^2 \le \sum_{i=1}^{N} p_i \mathbb{E}\|e_i^r\|^2 \le Me^2, \tag{82}$$

which proves Eq. (77). Notice that this worst-case bound requires neither unbiased compression nor independence among the client errors.

It remains to bound the global-gradient term appearing in $\mathcal{M}^r_{\text{dir}}$. At the beginning of round r, all clients are initialized at $w^r_{i,0} = w^r$. By the unbiasedness in Assumption C.2, conditional Jensen's inequality, and Assumption C.4,

$$\mathbb{E}\|g^r_{i,0}\|^2 = \mathbb{E}\|\nabla F_i(w^r)\|^2 + \mathbb{E}\|g^r_{i,0} - \nabla F_i(w^r)\|^2.$$

Because $0 \le \mathbb{E}\|g^r_{i,0} - \nabla F_i(w^r)\|^2$, and Assumption C.4

$$\mathbb{E}\|\nabla F_i(w^r)\|^2 \le \mathbb{E}\|g^r_{i,0}\|^2 \le \mathcal{G}^2. \tag{83}$$

Therefore,

$$\mathbb{E}\left\|\nabla F(w^r)\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{N} p_i \nabla F_i(w^r)\right\|^2 \le \sum_{i=1}^{N} p_i \mathbb{E}\left\|\nabla F_i(w^r)\right\|^2 \le \mathcal{G}^2 \tag{84}$$

Substituting Eq. (77) and Eq. (84) into the definition of $\mathcal{M}^r_{\text{dir}}$ gives Eq. (78).

For the simplified uniform bound, let $q = L\eta T \le 1/8$. The quantity inside the square root in Eq. (78) is at most

$$T^2\left[4\mathcal{G}^2 + 2\sigma_l^2 + 8q^2(2\mathcal{G}^2 + 2\sigma_g^2 + \sigma_l^2)\right] \le T^2\left(\frac{17}{4}\mathcal{G}^2 + \frac{1}{4}\sigma_g^2 + \frac{17}{8}\sigma_l^2\right), \tag{85}$$

which proves Eq. (79). Finally, both Eq. (77) and Eq. (79) hold uniformly over r. Summing them over R rounds therefore proves Eq. (80). □

D DATASET AND EXPERIMENTAL DETAILS

We implemented the models based on the PyTorch framework and ran the experiments with NVIDIA A5000 and A6000 processors. To smooth the convergence behavior in Figure 2, we use a rolling average with a window size of 11.

D.1 DETAILS OF PRIOR ALGORITHMS

FedPAQ (Reisizadeh et al., 2020): FedPAQ is a communication-efficient federated learning method based on Periodic Averaging and Quantization. In periodic averaging, a determined subset of clients participates in each communication round. In addition, participating clients upload quantized updates, further decreasing the communication payload. As a result, FedPAQ improves communication efficiency while maintaining learning performance.

SVDFed (Wang et al., 2023): SVDFed is a federated learning compression method that uses Singular Value Decomposition (SVD) to represent client gradients with a small number of basis vectors. Its main idea is that gradients from different clients often share correlated structures, which can be exploited through a low-rank approximation. Instead of transmitting full updates, clients communicate compressed representations reconstructed from these dominant singular components. This makes SVDFed an effective low-rank gradient compression approach for reducing communication overhead in FL.

MFedMC (Yuan et al., 2026): MFedMC is a communication-efficient multimodal federated learning (MFL) framework that reduces uplink cost by selecting only a subset of modalities to transmit in each round. Specifically, each client assigns a priority score to modalities and uploads only the encoder (or encoder update) of the selected modalities, thereby saving communication compared to transmitting all modality encoders. In MFedMC, the modality priority is computed using a weighted combination of (i) Shapley-based modality contribution, (ii) modality-wise model size, and (iii) the recency of modality encoder transmission.

MCB (Li et al., 2024): In MCB, each client selects modalities by considering both the modality contribution and the model size. Specifically, the contribution of modality m for client i at round r is computed using a confidence-weighted F1 score:

$$c_{i,m}^{r} = \sum_{j=1}^{K} w_j \, F1_{i,m}(j),$$

where w_j denotes the class weight and $F1_{i,m}(j)$ is the class-wise confidence-weighted F1 score. Based on this contribution and the model size $|\theta_{i,m}|$, the client-side utility is defined as

$$u_i^r(m) = \gamma \, \bar{c}_{i,m}^{r} + \delta \left(1 - \overline{|\theta_{i,m}|}\right),$$

where $\bar{c}_{i,m}^{r}$ and $\overline{|\theta_{i,m}|}$ are the normalized modality contribution and normalized model size, respectively, with $\gamma + \delta = 1$. Therefore, MCB prefers modalities that achieve higher confidence-weighted F1 scores while requiring smaller communication cost.

MFCPL (Le et al., 2025): MFCPL is a prototype-based MFL method that improves multimodal synergy by aligning (i) unimodal features and (ii) unimodal/fused features with *class prototypes* shared across clients. Let f_A and f_B denote unimodal features (for two modalities A and B) before fusion, and let f_M denote the fused feature.[4] To compare features in a common embedding space, MFCPL uses projection heads g_1 and g_2 that map fused and unimodal features to the same dimension: $z_M = g_1(f_M)$ and $z_m = g_2(f_m)$ for $m \in \{A, B\}$. It introduces three additional objectives:

1. **Cross-modal alignment (CMA).** It directly aligns unimodal features to encourage cross-modal consistency:
$$L_{\text{CMA}} := \|z_A - z_B\|_2^2.$$
2. **Cross-modal prototype contrastive (CMPC).** Let $P = \{p_k\}_{k=1}^{K}$ be the set of class prototypes and let p_y be the prototype of the ground-truth class y. MFCPL applies an InfoNCE-style contrastive loss to pull each unimodal embedding z_m toward its class prototype and

[4] Here, $f(\cdot)$ is used as a feature notation for convenience (distinct from the loss notation in the main paper).

push it away from other-class prototypes:

$$L_{\mathrm{CMPC}} := \sum_{m\in\{A,B\}} \left(-\log \frac{\exp\left(s(z_m, p_y)/\tau\right)}{\sum_{k=1}^{K} \exp\left(s(z_m, p_k)/\tau\right)} \right),$$

where $s(\cdot,\cdot)$ denotes cosine similarity and $\tau > 0$ is a temperature.

3. **Cross-modal prototype regularization (CMPR).** This term aligns the fused embedding with the (complete) class prototype, promoting global consistency across clients:

$$L_{\mathrm{CMPR}} := \|z_M - p_y\|_2^2.$$

Overall, MFCPL improves multimodal synergy by using prototype-based constraints that (i) align unimodal representations across modalities (CMA) and (ii) align unimodal/fused embeddings to class prototypes shared across clients (CMPC/CMPR).

D.2 Model Architectures

Table 4: Model Architectures of Dataset

Component	**CrisisMMD**		**KU-HAR**	**PTB-XL**	**CMU-MOSEI**
Modality	Image	Text	Acc & Gyro	I-to-aVF & V1-to-V6	Video (20) & Text (768) & Audio (5)
Encoder	**Linear**(1280, 128) **ReLU**() **Dropout**(p=0.1) **Linear**(128, 128)	**GRU**(768, 128, batch_first=True, dropout=0.1)	**Conv1d**(3, 32, kernel_size=5) **Conv1d**(32, 64, kernel_size=5) **Conv1d**(64, 128, kernel_size=5) **ReLU**() **MaxPool1d**(kernel_size=2) **Dropout**(p=0.1) **GRU**(128, 128, batch_first=True, dropout=0.1)	**Conv1d**(6, 32, kernel_size=5) **Conv1d**(32, 64, kernel_size=5) **Conv1d**(64, 128, kernel_size=5) **ReLU**() **MaxPool1d**(kernel_size=2) **Dropout**(p=0.1) **GRU**(128, 128, batch_first=True, dropout=0.1)	**Conv1d**(M, 32, kernel_size=5) **Conv1d**(32, 64, kernel_size=5) **Conv1d**(64, 128, kernel_size=5) **ReLU**() **MaxPool1d**(kernel_size=2) **Dropout**(p=0.1) **GRU**(128, 128, batch_first=True, dropout=0.1)
Fusion Network	**Linear**(128, 512) **Tanh**() **Linear**(512, 6)		**Linear**(128, 512) **Tanh**() **Linear**(512, 6)	**Linear**(128, 512) **Tanh**() **Linear**(512, 6)	**Linear**(128, 512) **Tanh**() **Linear**(512, 6)
Classifier	**Linear**(768, 64) **ReLU**() **Dropout**(p=0.1) **Linear**(64, 8)		**Linear**(768, 64) **ReLU**() **Dropout**(p=0.1) **Linear**(64, 8)	**Linear**(768, 64) **ReLU**() **Dropout**(p=0.1) **Linear**(64, 5)	**Linear**(768, 128) **ReLU**() **Dropout**(p=0.1) **Linear**(128, 3)

Our multimodal model consists of (i) a modality-specific encoder for each modality, (ii) an attention-based fusion module that produces a weighted fused representation, and (iii) a classifier for prediction. In Table 4, we summarize the model architectures for each dataset.

For **CrisisMMD**, each modality has a distinct encoder architecture, i.e., an MLP projection head for the image modality and a GRU-based encoder for the text modality. For **KU-HAR**, **PTB-XL**, and **CMU-MOSEI**, the modalities within each dataset share the same encoder design, consisting of a 1D convolutional encoder followed by a GRU (Acc/Gyro, I-to-aVF/V1-to-V6, and Text/Audio/Video, respectively). Specifically, **CMU-MOSEI** uses the same encoder architecture across modalities, while the feature dimensions differ: 20, 768, and 5 for video, text, and audio, respectively. Therefore, the modality-specific input dimension M in Table 4 is set to 20, 768, and 5 for the three modalities.

Across all datasets, the modality-specific representations are fused using an attention-based fusion module, and the resulting multimodal embedding is fed into a lightweight MLP classifier to produce the final predictions.

Preprocessing of each dataset: For CrisisMMD, we use pretrained backbones to extract raw-data features with MobileNet for images and DistilBERT for text, followed by encoders (MLP for image and GRU for text). For KU-HAR and PTB-XL, we encode normalized raw time-series signals using modality-specific encoders that stack Conv1d layers and a GRU, followed by an attention-based fusion module and an MLP classifier. For CMU-MOSEI, we use the preprocessed multimodal features provided by the MMSA framework (Yu et al., 2021), rather than processing the raw data directly. Specifically, we use BERT-based text features, FACET-based visual features, and COVAREP-based acoustic features.

D.3 HYPERPARAMETERS

Following the benchmark suggested in (Feng et al., 2023), we use the following common hyperparameter settings for all datasets: local learning rate as $\eta_l = 0.05$, communication rounds $R = 200$, and the number of local epochs is 1. For each dataset, we set the participation ratio of clients as 0.1 for **CrisisMMD** and **KU-HAR**, 0.25 for **PTB-XL**, and 0.1 for **CMU-MOSEI**. For the Softmax temperature τ, we tune over $\{10^{-1}, 1, 2, 5, 10\}$.

D.4 DATASETS

CrisisMMD (Alam et al., 2018) is a tweet-centric multimodal dataset that pairs user-posted images and text collected across seven major natural disasters in 2017 (earthquakes, floods, hurricanes, wildfires, etc.). It provides human annotations for (a) humanitarian relevance, (b) fine-grained humanitarian categories, and (c) damage severity. We follow prior work and formulate a disaster-response classification task. For FL partitioning, we distribute the data to 100 clients under an IID default following the benchmark (Feng et al., 2023).

KU-HAR (Sikder & Nahid, 2021) is the Human Activity Recognition (HAR) task dataset collected from 90 participants who are 18 to 34 years old on 18 different activities. It uses the accelerometer and gyroscope to predict the action. It is also partitioned by the person's ID. We split the total clients into 5 folds, with 63 training clients and the rest for validation and test data. Following the benchmark setting used in prior work (Feng et al., 2023), we consider eight activity classes: walking, walking upstairs, walking downstairs, sitting, standing, lying, jumping, and running.

PTB-XL (Wagner et al., 2022) is a large clinical ECG dataset with $21k$ ten-second, 12-lead recordings from $18k$ patients, annotated with up to 71 diagnostic statements that aggregate into five diagnostic superclasses (normal, myocardial infarction, ST/T change, conduction disturbance, hypertrophy). To construct a multimodal setting, we group leads into two modalities: I-AVF and V1-V6 following (Feng et al., 2023). The total number of hospital sites is 34, and each hospital is considered as the client where the patient's data is stored.

CMU-MOSEI (Bagher Zadeh et al., 2018) is a multimodal sentiment classification benchmark consisting of clipped samples from YouTube videos. Each sample contains text, audio, and video modalities, and is labeled as negative, neutral, or positive.

D.5 DISCUSSION ON MODALITY IMPORTANCE

Quantifying the exact importance of a factor in deep learning is inherently challenging because the final prediction arises from nonlinear interactions among multiple components.

For this reason, prior studies often rely on attribution methods that estimate the marginal contribution of each factor to model utility. *Shapley value* (Shapley, 1953) is a widely adopted and principled measure for this purpose, and has played a central role in evaluating the importance of various factors in machine learning, including data samples, model components, and participating clients in federated learning.

In multimodal learning, our factor of interest is the modality, i.e., how much each modality contributes to the final utility. To quantify this modality-wise contribution, we adopt the SHAPE score (Hu et al., 2022), which instantiates the Shapley-value framework in the multimodal setting by treating each modality as a player in the coalition game.

D.6 RESOURCE ALLOCATION AND COMMUNICATION COST

SVD implementation and rank assignment. We use modality-specific encoders based on the FedMultimodal benchmark (Feng et al., 2023). SVD compression is performed separately on the compressible weight matrices within each modality. A single modality-wise rank Λ_m is applied to all these matrices, rather than being divided among them. Let $\mathcal{L}_m$ denote the set of compressible weight matrices in modality m, and let $\ell \in \mathcal{L}_m$ index an individual weight matrix, denoted by $W_{m,\ell} \in \mathbb{R}^{p_{m,\ell} \times q_{m,\ell}}$. For each matrix, we transmit the retained left and right singular vectors and the corresponding singular values, using 32-bit precision for every transmitted scalar.

Correspondence between rank and communication budgets. The SVD configuration matches the encoder-wide transmission cost per unit rank across modalities within each dataset:

$$C_m := \sum_{\ell \in \mathcal{L}_m} (p_{m,\ell} + q_{m,\ell} + 1) = C.$$

When every matrix retains its assigned Λ_m components, the total SVD-factor payload per client is

$$B_{\text{SVD}} = 32 \sum_{m=1}^{M} \Lambda_m C_m = 32C \sum_{m=1}^{M} \Lambda_m \leq 32C\Lambda.$$

Thus, in this case, allocations with the same total retained rank incur the same SVD-factor communication cost, irrespective of how the ranks are distributed across modalities.

Implementation of modality-selection baselines. For MCB and MFedMC, we retain each method's modality-selection rule and apply the same SVD compressor to the selected modality's weight updates, while the selected modality receives the entire rank budget Λ. The assigned rank is applied to each compressed matrix of the selected modality, using the same 32-bit precision as `Fed-MIAC`. This setup compares binary modality selection with fine-grained modality-wise allocation using a common compression primitive.

Uncompressed parameters. Following prior SVD-based compression methods (Qiao et al., 2021; Wu et al., 2022), bias updates are not subject to low-rank compression. They are excluded from the rank-allocation problem because they have no associated SVD rank. Nevertheless, the reported encoder-upload costs include the transmitted bias updates in addition to the SVD factors.

E ADDITIONAL EXPERIMENT RESULTS

E.1 REAL-DEVICE EXPERIMENTS ON NVIDIA JETSON TX2

To evaluate the practicality of Fed-MIAC on real-world edge devices, we conduct additional experiments on NVIDIA Jetson TX2 platforms.

Hardware specifications: The NVIDIA Jetson TX2 platform is equipped with a 256-core NVIDIA Pascal GPU, a dual-core 64-bit NVIDIA Denver 2 CPU, and a quad-core ARM Cortex-A57 CPU.

Experimental setting: We measure the computation time on the devices and estimate the communication time by dividing the transmitted data size by an assumed uplink rate of 6 Mbit/s. This rate is chosen to represent a conservative Wi-Fi uplink condition based on IEEE 802.11g. The total execution time is estimated as the sum of the measured computation time and the estimated communication time.

Compared with the simulation environment, the limited computing capabilities of edge devices result in longer execution times for SVD, SHAPE score computation, and model training. Nevertheless, Fed-MIAC consistently achieves a shorter estimated total execution time than the baselines owing to its reduced communication overhead, exhibiting trends consistent with those observed in the simulation results.

E.2 APPLYING MODALITY IMPORTANCE TO QUANTIZATION-BASED COMPRESSION

Table 5: Test F1-score of FedPAQ with modality importance under 6- and 4-bit quantization.

Dataset	**CrisisMMD**			**KU-HAR**			**PTB-XL**		
Quant. Bits	Full	6	4	Full	6	4	Full	6	4
FedPAQ	50.30	43.91	33.63	66.74	61.88	48.07	61.83	60.29	54.56
+MIAC		**48.90**	**44.64**		**63.21**	**61.45**		**61.86**	**61.22**

To further verify the effectiveness of modality-aware resource allocation, we apply our modality-importance-aware strategy (denoted as +MIAC) to a basic quantization-based compression method, FedPAQ (Reisizadeh et al., 2020). We exclude this comparison from the main table because the quantization levels cannot be exactly matched to the SVD rank settings. To ensure a comparable compression rate between quantization and SVD, we set the total quantization bit budgets to 6 and 4, corresponding to uniform per-modality allocations of $(3, 3)$ and $(2, 2)$ bits, respectively. These settings yield compressed model sizes comparable to those obtained with dataset-specific SVD ranks of $(32, 16, 16)$ and $(16, 8, 8)$ for CrisisMMD, KU-HAR, and PTB-XL, respectively.

Table 5 reports the results, where **Quant. Bits** denotes the total quantization bit budget across modalities. For reference, we also report the uncompressed (**Full**) performance shown in the main results (Table 1). FedPAQ allocates the bit budget uniformly across modalities, whereas +MIAC allocates more bits to more important modalities. +MIAC consistently improves performance across all datasets and bit budgets, with particularly large gains under 4-bit quantization, ranging from $+6.66$ to $+13.38$. These results show that the effectiveness of incorporating modality importance into compression is not tied to a specific compression scheme.

E.3 CONFIDENCE INTERVAL OF FL ALGORITHMS

We report the 95% confidence intervals corresponding to Table 1. We run CrisisMMD and PTB-XL over five seeds, but ten seeds for KU-HAR to improve the reliability of the results, because KU-HAR shows larger confidence intervals with five seeds. CrisisMMD and PTB-XL datasets exhibit relatively tight intervals across rank budgets, indicating stable performance across runs. In contrast, KU-HAR shows larger variability across seeds, suggesting that its performance is more sensitive to stochastic factors (e.g., random initialization or client participation) under our federated setting. Nevertheless, the mean performance consistently favors Fed-MIAC over baselines across rank budgets, indicating a robust advantage in expectation despite higher variance on KU-HAR.

Table 6: 95% confidence interval of FL algorithms on the CrisisMMD, KU-HAR, and PTB-XL datasets

Base FL	Method	CrisisMMD					KU-HAR					PTB-XL				
	Λ (Rank)	Full	32	16	8	4	Full	32	16	8	4	Full	32	16	8	4
FedAvg	SVDFed	1.50	3.21				8.42	6.59				0.69	1.35			
	MCB		1.65	0.45	2.90	1.82		4.49	5.34	7.81	5.21		1.08	1.02	2.02	2.33
	MFedMC		1.04	1.19	3.06	1.19		5.99	5.20	4.97	5.64		0.97	1.69	1.30	1.43
	Fed-MIAC		1.11	2.08	2.32	3.14		7.30	7.41	5.06	8.16		0.72	0.67	0.58	2.16
MFCPL	SVDFed	1.50	3.28				6.56	7.16				0.76	2.91			
	MCB		1.08	1.62	3.84	2.31		5.65	7.03	6.51	5.85		0.72	0.58	0.95	1.20
	MFedMC		2.41	2.59	2.24	1.52		6.32	5.07	5.26	4.97		1.48	1.89	2.00	1.56
	Fed-MIAC		1.33	1.61	2.87	2.22		6.86	7.54	5.91	7.30		0.98	1.54	1.70	1.18

E.4 Experimental results on CMU-MOSEI dataset

Base FL	Method	CMU-MOSEI				
	Λ (**Rank**)	Full	32	16	8	4
FedAvg	SVDFed	60.84	49.35			
	MCB		54.55	54.55	52.45	52.32
	MFedMC		42.84	42.21	40.50	40.37
	Fed-MIAC		**61.10**	**60.83**	**60.07**	**59.81**
MFCPL	SVDFed	62.40	59.54			
	MCB		61.34	61.82	61.51	61.64
	MFedMC		54.88	58.75	53.78	50.39
	Fed-MIAC		**62.65**	**62.64**	**62.26**	**61.91**

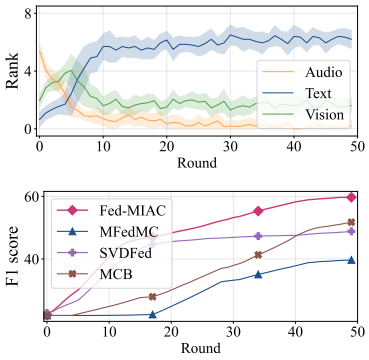

Figure 4: Experimental results on the CMU-MOSEI dataset: **(Left)** test F1-score, **(Upper-Right)** behavior of rank allocation, and **(Lower-Right)** FL convergence.

We evaluate the CMU-MOSEI dataset across three modalities: audio, text, and vision for sentiment classification, as shown in Figure 4.

The table in Figure 4 **(Left)** shows that Fed-MIAC consistently outperforms the other baselines across all compression levels under both FedAvg and MFCPL, demonstrating that adaptive rank allocation remains effective as the number of modalities increases to three. As shown in Figure 4 **(Upper-Right)**, Fed-MIAC consistently allocates a larger rank to text while assigning smaller ranks to vision and audio, enabling it to preserve information from all modalities while concentrating the communication budget on the more important ones.

In contrast, modality-selection-based methods become more challenging as the number of modalities increases. In the three-modality setting, selecting only a subset of modalities at each round reduces the transmission opportunities for the remaining modalities. This issue can be further amplified in MFedMC, whose recency-based criterion favors modalities that have not been selected for a longer period, potentially prioritizing recency over their current contribution to the task. Fed-MIAC, on the other hand, avoids hard modality selection by continuously adjusting the compression level of each modality.

Consequently, as shown in Figure 4 **(Lower-Right)**, Fed-MIAC exhibits substantially more stable convergence than the selection-based baseline and achieves performance close to the Full setting across different compression levels.

E.5 EXTENSION OF Fed-MIAC TO LARGE MODELS

Table 7: Computation and uplink time of LoRA and Fed-MIAC.

Method	Comput. Ratios	Comput. Time (s)	Uplink Time (s)	Total Time (s)
LoRA	–	74.69	5,125.67	5,200.36
Fed-MIAC	[36.60%, 23.84%, 39.56%]	118.49	3,552.63	3,671.12 (**29.41%**↓)

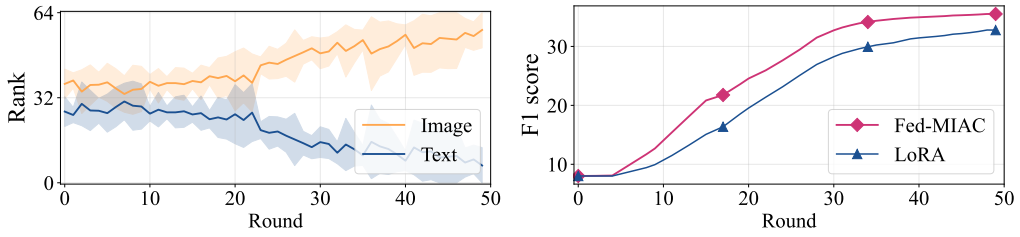

Figure 5: Behavior of rank allocation (Left) and FL convergence (Right) on the CrisisMMD dataset.

To evaluate Fed-MIAC in a large-model setting, we compare it with LoRA under approximately matched communication budgets. We employ pretrained ViT-Base/32 and BERT-small as the image and text encoders, respectively, and train on raw RGB images and tweet text without feature extraction, whereas CrisisMMD in the main paper uses extracted features. For preprocessing, we perform only image resizing and normalization for the image modality and tokenization for the text modality.

For LoRA, rank-r adapters are inserted into the query and value projection layers of every self-attention layer in ViT and BERT. The pretrained backbone weights remain frozen, and only the learned adapter parameters are transmitted to the server without additional compression. For Fed-MIAC, we directly optimize the original dense query and value projection weights. Each client computes the resulting model updates and dynamically allocates a shared TSVD rank budget across the image and text modalities according to their SHAPE-based modality importance.

We set LoRA rank $r = 32$ and Fed-MIAC budget $\Lambda = 64$. This setting accounts for the fact that LoRA applies a rank-r update separately to each modality, whereas Fed-MIAC distributes a single shared rank budget across the two modality-specific updates. The resulting payload sizes are not exactly identical because LoRA and TSVD use different parameterizations, with TSVD additionally transmitting the retained singular values, whose overhead is negligible.

Table 7 reports the cumulative computation and uplink time required for each method to first reach Macro-F1$= 33$, corresponding to the performance of LoRA at Round 50. Although Fed-MIAC incurs additional computation for SHAPE-score estimation and SVD, its practical efficiency should be assessed by the total time including both computation and communication. Within the computation time of Fed-MIAC, local training, SVD, and SHAPE-score estimation account for 36.60%, 23.84%, and 39.56%, respectively. Compared with LoRA at rank 32, Fed-MIAC with $\Lambda = 64$ incurs an additional 43.80 seconds of computation while reducing the uplink time by 1,573.04 seconds, resulting in a 29.41% reduction in total time. These results show that the communication savings outweigh the additional SHAPE-score and SVD computation under the bandwidth-constrained setting.

Beyond the wall-clock comparison, we further analyze how Fed-MIAC distributes its shared rank budget across modalities. The left panel of Figure 5 shows that Fed-MIAC progressively allocates more rank to the image modality while reducing the rank assigned to the text modality, consistent with the modality-dependent allocation observed in the main experiments. The right panel shows that Fed-MIAC converges faster than LoRA and achieves a higher Macro-F1 score at Round 50 (35.50 vs. 33.52). Taken together, the rank-allocation, convergence, and wall-clock results demonstrate that the communication-efficiency benefit of Fed-MIAC extends to the large-model setting.

E.6 Effect of modality importance in rank allocation

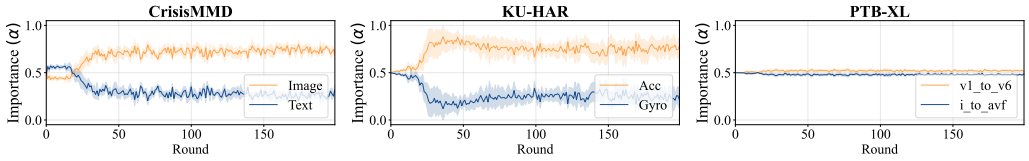

Figure 6: Behavior of normalized modality importance (SHAPE score) across the three datasets.

To further investigate modality importance, we plot the normalized SHAPE score across datasets in Figure 6. The score is measured at rank budget 8 and averaged over seeds and participating clients at each round.

CrisisMMD and KU-HAR exhibit clearly imbalanced modality importance, with the image and accelerometer modalities playing dominant roles, respectively, whereas PTB-XL shows a more balanced importance. Although the modality importance in PTB-XL is relatively balanced, incorporating modality importance still leads to performance gains, as shown in Table 3.

These results in Table 3 and Figure 6 suggest that incorporating modality importance into rank allocation is important for communication-efficient MFL and support our claim that it should be taken into account when allocating compression budgets.

E.7 Sensitivity analysis of softmax temperature τ

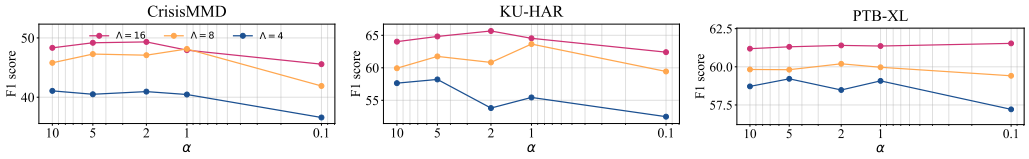

Figure 7: Impacts of parameter τ across the three datasets

Figure 7 analyzes the impact of the softmax temperature τ used to normalize the SHAPE score, which is the Shapley-based modality importance weights. Overall, `Fed-MIAC` maintains strong performance across a wide range of τ values and consistently outperforms the baseline. We study the sensitivity to the softmax temperature τ across ranks 4, 8, and 16. Overall, the method achieves consistently strong performance over a broad range of values, with $\tau \in [2, 5]$ often yielding the best or near-best results. This indicates that the method is reasonably robust to the choice of τ.

Across datasets, performance decreases when τ is small, i.e., 0.1. Since we use $\mathrm{softmax}(s/\tau)$, decreasing τ makes the importance distribution increasingly sharp and causes the weights to concentrate on only a few components. Consequently, the importance-based weights can dominate the contribution of the singular values, effectively diminishing the role of singular-value magnitudes in SVD-based compression. This excessive concentration may discard complementary information carried by other components. In contrast, moderate values of τ preserve the relative ordering of importance scores while providing a smoother allocation of weights across components. These results suggest that moderate softmax temperatures offer a robust balance between emphasizing important components and retaining the information encoded by singular values.

E.8 RANK ALLOCATION UNDER HETEROGENEOUS BANDWIDTH

Table 8: Ablation on heterogeneous rank allocation

Method	CrisisMMD	KU-HAR	PTB-XL
Fixed Heterogeneous Rank	47.91	63.63	60.30
Resampled Heterogeneous Rank	49.43	62.97	60.73
Fixed Uniform Rank in Main Table 1	48.14 ($\Lambda = 8$)	63.65 ($\Lambda = 8$)	60.20 ($\Lambda = 8$)

In practical edge-network environments, uplink bandwidth can vary across clients and may also fluctuate over communication rounds. Therefore, enforcing an identical rank budget for all clients may be restrictive, especially when clients have heterogeneous communication capacities.

To examine the robustness of `Fed-MIAC` under such conditions, we consider two heterogeneous rank-allocation settings. In Fixed Heterogeneous Rank, each client is assigned a different rank sampled from $\{4, 8, 16, 32\}$, and the assigned rank remains fixed throughout training. In Resampled Heterogeneous Rank, each client samples its rank from the same set at every communication round, mimicking time-varying communication conditions.

As shown in Table 8, `Fed-MIAC` remains effective under both heterogeneous settings. Compared with the fixed uniform-rank setting used in Table 1, where all clients share the same fixed rank Λ throughout training, heterogeneous rank allocation achieves comparable or robust performance on all three datasets. These results demonstrate that Fed-MIAC accommodates client-specific and time-varying rank budgets.

E.9 LESS FREQUENT SHAPE SCORE FOR COMPUTATIONAL COST

Table 9: Ablation on SHAPE score recomputation frequency

Method	CrisisMMD		KU-HAR		PTB-XL	
Λ **(Rank)**	8	4	8	4	8	4
`Fed-MIAC` ($N_s = 1$)	48.14	41.07	63.65	58.21	60.20	59.21
`Fed-MIAC` ($N_s = 10$)	45.81	40.90	62.67	57.38	60.09	58.66
MCB	43.72	37.63	56.35	49.76	57.03	55.65
MFedMC	41.55	37.43	58.14	54.96	55.57	55.42

While our primary goal is to reduce *communication cost*, SHAPE score incurs additional computation, i.e., 2^M coalition evaluations. We investigate the effect of reducing the computation overhead of Shapley-based modality importance by recomputing the SHAPE score less frequently in Table 9. Here, N_s denotes the participation-based recomputation period: each client recomputes its SHAPE score once every N_s local participations. Thus, $N_s = 1$ corresponds to our default setting used in the main experiments, where a client recomputes the SHAPE score whenever it participates. In contrast, $N_s = 10$ recomputes the SHAPE score every 10 participations, reducing the SHAPE score recomputation frequency (and its associated overhead) by roughly $10\times$. As shown in Table 9, using a larger N_s leads to a mild performance drop compared to $N_s = 1$. Nevertheless, even with $N_s = 10$, `Fed-MIAC` remains consistently better than **MFedMC** and **MCB** across datasets and budgets, suggesting that the SHAPE score is sufficiently stable to retain most of the benefits even under infrequent recomputation.