Towards Optimal Multi-Modal Federated Learning on Non-IID Data with Hierarchical Gradient Blending

Sijia Chen
Department of Electrical and Computer Engineering
University of Toronto
sjia.chen@mail.utoronto.ca

Baochun Li
Department of Electrical and Computer Engineering
University of Toronto
bli@ece.toronto.edu

IEEE INFOCOM 2022 - IEEE Conference on Computer Communications | 978-1-6654-5822-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INFOCOM48880.2022.9796724

***Abstract*—Recent advances in federated learning (FL) made it feasible to train a machine learning model across multiple clients, even with non-IID data distributions. In contrast to these *uni-modal* models that have been studied extensively in the literature, there are few in-depth studies on how *multi-modal* models can be trained effectively with federated learning. Unfortunately, we empirically observed a counter-intuitive phenomenon that, compared with its uni-modal counterpart, multi-modal FL leads to a significant degradation in performance.**

Our in-depth analysis of such a phenomenon shows that modality sub-networks and local models can overfit and generalize at different rates. To alleviate these inconsistencies in collaborative learning, we propose hierarchical gradient blending (HGB), which simultaneously computes the optimal blending of modalities and the optimal weighting of local models by adaptively measuring their overfitting and generalization behaviors. When HGB is applied, we present a few important theoretical insights and convergence guarantees for convex and smooth functions, and evaluate its performance in multi-modal FL. Our experimental results on an extensive array of non-IID multi-modal data have demonstrated that HGB is not only able to outperform the best uni-modal baselines but also to achieve superior accuracy and convergence speed as compared to state-of-the-art frameworks.

***Index Terms*—multi-modal machine learning, federated learning, gradient blending**

I. Introduction

In recent years, research interests in training a machine learning model on edge clients motivated the paradigm of federated learning (FL) [1]–[3], which makes it feasible for multiple devices to collaboratively train a global model in a privacy-preserving manner. Existing federated learning mechanisms [4]–[6] focused on training *uni-modal* models such as image classifiers and text predictors, either assuming that data is independent and identically distributed (IID) across the clients, or with more challenging non-IID data distributions [7]–[9].

Compared with uni-modal models, *multi-modal* models [10] with multiple input modalities has become a recent trend in machine learning. Though several existing studies [11], [12] proposed to train multi-modal models with federated learning, the specific challenges imposed by non-IID data distributions in the context of multi-model federated learning have not been explored.

In this paper, we foray into uncharted territory by focusing on the effects of non-IID data distributions in the context of multi-modal federated learning. Intuitively, as more information is available from complementary modalities, multi-modal FL should outperform the corresponding uni-modal FL. However, we make the counter-intuitive observation from our experimental results that multi-modal FL not only leads to more communication rounds needed to reach convergence, but also lags far behind conventional uni-modal FL with respect to the accuracy of trained models.

In this paper, we discovered the root cause for such performance degradation. Due to the complexity of non-IID multi-modal data, both the local data distribution across clients and the distribution between modalities can be extremely heterogeneous. As a result, multi-modal models are often prone to overfitting as they are trained on unbalanced and small-size local datasets. To make matters worse, overfitting and inconsistent generalization rates appear in the modality sub-networks and the local models simultaneously. With uni-modal FL, existing mechanisms have been proposed to address the problems of overfitting and heterogeneity among clients, such as local adaptation [13]–[15] and weighted global aggregation [16], [17]. However, our experiments show that these mechanisms were not able to provide an effective solution in the context of multi-modal FL.

We present a thorough theoretical analysis to quantitatively analyze *weight divergence*, which represents the difference between weights updated based on non-IID multi-modal data and centralized data. Our analysis shows that inconsistencies found in each training stage can increase the overall divergence. Built upon these theoretical insights, the crux of this paper is a new mechanism, referred to as *hierarchical gradient blending* (HGB), that adaptively computes an optimal blending of modalities and reweigh updates from the clients according to their overfitting and generalization behaviors. Intuitively, HGB corresponds to an optimization problem with overfitting-to-generalization rate minimization as the objective function. The obtained optimal weights can reduce generalization error

when training the global model. Our new mechanism does not introduce any trainable parameters, making it computationally friendly and easy to use. From a theoretical perspective, we show that HGB guarantees convergence in multi-modal FL with non-IID data.

Our original contributions in this paper are as follows. *First*, to our best knowledge, our work is the first to address the performance challenges when multi-modal FL is used with non-IID data distributions. *Second*, with rigorous theoretical analyses, we demonstrate the root cause for such performance degredation, in that both local updates and the global aggregation suffer from overfitting and inconsistent generalization rates. *Third*, we propose a new mechanism, hierarchical gradient blending (HGB), which adaptively achieves the optimal blending of modalities and the optimal aggregation of clients' updates. *Finally*, with several benchmark multi-modal datasets, we evaluate HGB's performance in the context of multi-modal FL experimentally, and show that it outperforms state-of-the-art FL mechanisms by a substantial margin in terms of both accuracy and the speed of convergence.

II. Related Work

Uni-modal federated learning. Existing mechanisms in federated learning focused on training a shared global model effectively, with the hope that its performance is competitive with models trained with centralized data. However, existing mechanisms were limited to training a uni-modal model, which utilizes one modality as input. As examples, *uni-modal* FL mechanisms were solely proposed to train an effective image classifier or text predictor, which contains a single network of one modality, and based on independent and identically distributed (IID) and non-IID data across the clients.

Multi-modal federated learning. As multiple modalities can provide more information, multi-modal models outperformed uni-modal models in many applications, such as video classification. There are a very limited amount of existing work in the literature proposing to train multi-modal models in FL settings. Liu *et al.* [11] generated a powerful representation from multiple task-oriented representations obtained by the federated learning framework. Liang *et al.* [12] evaluated its proposed FL method on a task related to Visual Question Answering (VQA). However, in-depth analysis and discussions on challenges in multi-modal FL are still missing.

Our work in this paper attempts to present such an in-depth analysis in the context of late-fusion multi-modal FL, in which multi-modal models have a late-fusion structure that combines the outputs of sub-networks for different modalities to make the prediction. The capacity and complexity of a multi-modal model are significantly higher than its uni-modal counterpart.

Client heterogeneity. One of the critical challenges in FL is heterogeneity across participating clients. Local adaptation [13]–[15], [18] aimed to train a global model that can generalize well on each device's local data. Ji *et al.* proposed FedAtt [8], [16], which aggregates model updates from the updates with biased weights in order to train models with higher qualities. [19], [20] jointly optimized mixed global and local models to seek a trade-off between overfitting and generalization. However, in a range of non-IID multi-modal data scenarios, we will show in this paper that existing FL mechanisms were not able to train effective multi-modal models, or to reach convergence within an acceptable number of communication rounds.

Our work is inspired by Wang *et al.* [21], which observed the inherent overfitting problem when training a late-fusion multi-modal model. The gradient-blending schema used in the literature [21]–[23] serves as the foundation for our proposed algorithm, *hierarchical gradient blending*. In essence, our work is mostly related to current works that assigned weights and probabilities to clients or added auxiliary regularization terms to the learning objective. While all existing mechanisms relied upon the ability to learn weights during training and hope to achieve a balanced trade-off between overfitting and generalization, our work is both task-agnostic and architecture-agnostic, and optimizes weights directly without additional learning in the context of multi-modal FL.

III. Preliminaries

Federated learning aims to train a classification model $\boldsymbol{w}$ based on the dataset D that is separately stored in C clients in which each client c_k contains its own train set $D^k : \{\boldsymbol{s} : (\boldsymbol{X}, y)\}$ and evaluation set $D'^k : \{\boldsymbol{s}' : (\boldsymbol{X}', y')\}$ where the sample $\boldsymbol{X}$ and the true label y of D^k, D'^k are sampled from the local data distribution $(\mathcal{X}, \mathcal{Y})^k$. FL considers the following distributed optimization problem:

$$\min_{\boldsymbol{w}} \left\{ F(\boldsymbol{w}) = \sum_{k=1}^{C} p_k f_k(\boldsymbol{w}) \right\} \tag{1}$$

where $p_k \geq 0$ is the weight of k-th client and the objective function f_k of k-th client follows assumptions 1 and 2.

Assumption 1 (Convex and Smooth)**.** *The objective functions $f_1, ..., f_C$ of clients are all convex and L-smooth.*

Assumption 2 (Smoothness property)**.** *For any function $f \in [f_1, ..., f_C]$, the smoothness property implies the inequality as $\| \nabla f(y) - \nabla f(y) \| \leq L \| y - x \|$.*

A. Uni-modal Federated Learning

The conventional uni-modal FL considers a uni-modal model that contains a deep network ψ_m with parameters $\boldsymbol{v_m}$. ψ_m utilizes a single modality $\boldsymbol{X} = \boldsymbol{x_m}$ (e.g. RGB frames, audio, or optical flows) as input for further classification. Thus, the f_k is computed to minimizing an empirical loss:

$$f_k = \frac{1}{|D^k|} \sum_{s \in D^k} l\left(\psi_m(\boldsymbol{x_m}; \boldsymbol{v_m}); y\right) \tag{2}$$

where l is a user specific loss function, such as cross entropy.

Minimizing Eq. (1) of the uni-modal FL gives the solution $\boldsymbol{w}^* = \boldsymbol{v_m^*}$ for the network of modality m.

B. The Distributed Training Process

Our paper uses the distributed mini batch stochastic gradient descent (MB-SGD) algorithm to solve Eq. (1). In the FL setting, the training paradigm proposed by FedAvg [1] is constructed in three stages, including local update, global aggregation, and aggregated model broadcasting.

Let $\boldsymbol{w}$ denote parameters that are known to both the server and the client, and $\boldsymbol{v}$ denote parameters that are only visible to the client. The number of local update steps is E. During the total T training iterations with the iteration index $t = 0, 1, ..., T$, the number of communication rounds is $P = \frac{2T}{E}$ with the index $p = 0, 1, ..., P$. Then, in the p-th global aggregation, the local parameter $\boldsymbol{v}_t^k$ of the client k is updated by the MB-SGD where t ranges from t_{p-1} to $t_p - 1$. At $t = 0$, the local parameters for all clients are initialized to the $\boldsymbol{w}_0$.

$$\begin{aligned}
\text{local update}: &\begin{cases} \boldsymbol{v}_{t+1}^k = \boldsymbol{v}_t^k - \eta_k \widetilde{g}(\boldsymbol{v}_t^k) & t+1 \notin \mathcal{I}_e \end{cases} \\
\text{global aggregation}: &\begin{cases} \boldsymbol{w}_{t+1}^k = \boldsymbol{v}_{t+1}^k \\ \boldsymbol{w}_{t+1} = \sum_{k=1}^K p_k \boldsymbol{w}_{t+1}^k & t+1 \in \mathcal{I}_e \end{cases} \\
\text{broadcast}: &\begin{cases} \boldsymbol{v}_{t+1}^k = \boldsymbol{w}_{t+1} \end{cases}
\end{aligned} \tag{3}$$

where $\mathcal{I}_e = \{t_0, t_1, ..., t_P\}$ with $E = t_{p+1} - t_p$. $\widetilde{g}_t^k := \widetilde{g}(\boldsymbol{v}_t^k) = \nabla f_k(\boldsymbol{v}_t^k)$ is the "true" gradient computed with the target local data distribution. In MB-SGD, we compute the $g_t^k := g(\boldsymbol{v}_t^k; \xi^k) = \frac{1}{\xi^k} \sum_{s \in \xi^k} \nabla f_k(\boldsymbol{v}_t^k, s)$ with a mini-batch of samples ξ^k drawn from the local distribution of client k. $\widetilde{g}_t^k$ and g_t^k follows the unbiased assumption 3.

Assumption 3 (Unbiased gradients)**.** $E_{\xi^k}[g(\boldsymbol{v}_t^k; \xi^k)] = \widetilde{g}(\boldsymbol{v}_t^k)$

C. Overfitting-to-generalization Rate

The overfitting-to-generalization rate (OGR) is a metric that evaluates the performance of the trained model on both training and validation datasets. It is defined as follows:

$$\begin{aligned}
OGR(n_1, n_2) &= \frac{\triangle O(n_1, n_2)}{\triangle G(n_1, n_2)} \\
&= \frac{\mid \left(l^T(\boldsymbol{w}_{n_1}) - l^T(\boldsymbol{w}_{n_2})\right) - \left(l^*(\boldsymbol{w}_{n_1}) - l^*(\boldsymbol{w}_{n_2})\right) \mid}{\mid l^*(\boldsymbol{w}_{n_1}) - l^*(\boldsymbol{w}_{n_2}) \mid}
\end{aligned} \tag{4}$$

where $n_1 < n_2$ is the train step index. l^T is the loss in the train set while l^* is the ground-truth loss computed in the target distribution.

Based on MB-SGD, in an adjacent parameter update step $\boldsymbol{w}_{n_2} = \boldsymbol{w}_{n_1} - \eta g$ with a small gradient g, the $l^T(\boldsymbol{w}_{n_2})$ and $l^*(\boldsymbol{w}_{n_2})$ can be extended by Taylor theorem. Then, OGR becomes:

$$OGR(n_1, n_2) = \frac{\langle \nabla l^T - \nabla l^*, g \rangle}{\langle \nabla l^*, g \rangle} \tag{5}$$

D. The Gradient Blending Schema

Typically, gradient blending (GB) is a linear combination of gradients from different models. The works of multi-task generally utilize GB to train models of different tasks to achieve joint optimization. As each gradient is direct derived from the loss function, GB re-weights loss functions to obtain a blended auxiliary loss shown as $L_{blend} = \sum_{m=1}^M z_m l_m$ where l_m is the m-th loss while z_m is the corresponding weight. However, the blending weights need to be initialized and set manually.

Our work was initially motivated by the *optimal gradient blending* proposed in the paper [21]. It computed the best weights $\{z_m\}_{m=1}^M$ by minimizing a measurable OGR metric. The direct benefit of this schema is that the computed z_m^* induces better generalization behavior without additional learning processes.

IV. Multi-Modal Federated Learning

The late-fusion multi-modal FL considers the training of models $\{\boldsymbol{v_m}\}_{m=1}^M$ with M modalities as inputs. Let each sample used in training be $\boldsymbol{X} = \{\boldsymbol{x_m}\}_{m=1}^M$. Each modality m has its own deep network ψ_m with parameter $\boldsymbol{v_m}$. These sub-networks are jointly trained to minimize the empirical loss:

$$f_k = \frac{1}{|D^k|} \sum_{s \in D^k} l\left(\mathbb{C}(\psi_1(\boldsymbol{x_1}), ..., \psi_M(\boldsymbol{x_M})); y\right) \tag{6}$$

where $\mathbb{C}$ is a classifier used to process the outputs of M sub-networks to make the prediction. $\mathbb{C}$ can either be designed as a fusion operation followed by fully-connected (FC) layers or directly average the prediction scores from M sub-networks.

A. Non-IID Multi-modal Data in FL

Compared with uni-modal data $\boldsymbol{x_m}$ generated from one distribution, multi-modal data $\{\boldsymbol{x_m}\}_{m=1}^M$ is sampled from the joint distribution of M modalities (i.e., RGB frames, audio, and optical flow). This leads to both lable-based non-IID and the modality-based non-IID. Besides, one client may contain either only part of modalities or samples with incomplete modalities. Therefore, *the gradient distance between any two clients* $||\nabla f_k(\boldsymbol{v}_t^k; \xi^k) - \nabla f_n(\boldsymbol{v}_t^n; \xi^n)||$ *cannot be bounded by a finite (fixed) constant.*. This further induces that $||\nabla f_k(\boldsymbol{v}_t^k; \xi^k)||$ is not bounded. This leads to our proposition 1 on the stochastic gradient of each client.

Proposition 1 (Unbounded stochastic gradient)**.** *The gradient discrepancy* $\parallel \widetilde{g}(\boldsymbol{v}_t^i) - \widetilde{g}(\boldsymbol{v}_t^j) \parallel^2$ *between clients* i *and* j *cannot be bounded by a finite (fixed) constant. This leads to the stochastic gradients* $E_{\xi^k}||\nabla f_k(\boldsymbol{v}_t^k; \xi^k)||^2$ *in client* k *is not uniformly bounded in the non-IID multi-modal data.*

Proof: We have $||\nabla f_k(\boldsymbol{v}_t^k; \xi^k)||^2 = ||E_{\xi^k}[\nabla f_k(\boldsymbol{v}_t^k; \xi^k)]||^2$ that is not bounded. Then, as $||E_{\xi^k}[\nabla f_k(\boldsymbol{v}_t^k; \xi^k)]||^2 \leq E_{\xi^k}||\nabla f_k(\boldsymbol{v}_t^k; \xi^k)||^2$, we obtain that $E_{\xi^k}||g(\boldsymbol{v}_t^k; \xi^k)||^2$ is not uniformly bounded. ■

This further induces a new bound for the gradient variance of participating clients, shown by lemma 1.

Lemma 1 (Bounding the gradient variance)**.** *For* K *participating clients, the expected gradient is* $\widetilde{\overline{g_t}} = \sum_{k=1}^K \widetilde{g}(\boldsymbol{v}_t^k)$ *while the computed stochastic gradient is* $\overline{g_t} = \sum_{k=1}^K g(\boldsymbol{v}_t^k)$*. The upper bound for gradient variance follows:*

$E \parallel \widetilde{\overline{g_t}} - \overline{g_t} \parallel^2 \leq \sum_{k=1}^K p_k^2 \left[4L\left(f_k(\boldsymbol{v}_t^k) - f_k(\boldsymbol{v}_*^k)\right) + 2E \parallel \nabla f_k\left(\boldsymbol{v}_*^k; \xi\right) \parallel^2\right]$
where $\boldsymbol{v}_*^k$ *is the optimal value of the* L*-smooth function* f_k *based on the local dataset* D_k*.*

Proof: According to $(\sum_{n=1}^{N} a_n)^2 = \sum_n a_n^2 + 2\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} a_i a_j$ and Assumption 2, the gradient variance becomes $E \parallel \widetilde{\overline{g_i}} - \overline{g_i} \parallel^2 = \eta_i^2 \sum_k p_k^2 E||(\nabla f_k(\boldsymbol{v}_t^k) - \nabla f_k(\boldsymbol{v}_t^k, \xi_t^k)||^2]$. Then, based on $E \parallel X - EX \parallel^2 \leq E \parallel X \parallel^2$ and the $\frac{1}{2} \parallel a \parallel^2 - \parallel b \parallel^2 \leq \parallel a - b \parallel^2$, we can complete the proof by obtaining $E \parallel \nabla f_k(\boldsymbol{v}_t^k, \xi_t^k) \parallel^2 \leq 4L \left(f_k\left(\boldsymbol{v}_t^k\right) - f_k\left(\boldsymbol{v}_*^k\right)\right) + 2E \parallel \nabla f_k(\boldsymbol{v}_*^k, \xi) \parallel^2$. ■

B. Challenges

The non-IID multi-modal data can induce overfitting and inconsistent generalization rates between sub-networks of modalities and between local models of clients. For the former, the multi-modal model that contains M sub-networks $\{\psi_m\}_{m=1}^{M}$ for modalities is prone to overfitting because of updating the model with large-scale parameters based on limited-size local data. Then, as pointed by work [21], these sub-networks with different structures naturally contain different inherent generalization rates. Besides, the incomplete modalities of local samples further introduce gradients divergence when jointly training these sub-networks. For the latter, the gradient variance is oriented by the diversity degree among clients, as shown in lemma 1. Thus, starting from the shared global model, the local models of different clients can generalize at extremely different rates.

The problem of the non-IID multi-modal data can be presented by the weight divergence $\parallel \boldsymbol{w}_{t_p}^{(f)} - \boldsymbol{w}_{t_p}^{(c)} \parallel$, as shown in work [7], where $\boldsymbol{w}_{t_p}^{(f)}$ and $\boldsymbol{w}_{t_p}^{(c)}$ are the weights of t_p-th round in the FL training and centralized training, respectively. We denote the multi-modal cross-entropy loss as:

$$\nabla \ell(\{\boldsymbol{v_m}\}_{m=1}^{M}) = \sum_{m=1}^{M} z_m^k \sum_{i=1}^{|\mathcal{Y}|} p_m(y=i) \nabla \psi_{(i)}(\boldsymbol{x_m}; \boldsymbol{v_m}) \tag{7}$$

where $\nabla \psi_{(i)}(\boldsymbol{x_m}; \boldsymbol{v_m}) = \nabla E_{\boldsymbol{x_m}/y_m=i}[\log \psi_{(i)}(\boldsymbol{x_m}, \boldsymbol{v_m})]$ is L_{mi}-Lipschitz and z_m^k is the weight of modality m in client k.

Local divergence of j-th step update in each client k can be computed as the gradient distance between $\boldsymbol{v}_j^k$ and $\boldsymbol{w}_j$, i.e., $\parallel \nabla f_k(\boldsymbol{v}_j^k, \xi^k) - \nabla f_k(\boldsymbol{w}_j, \xi^k) \parallel$. Then, applying Eq. (7) to this distance, we can get its upper bound as $\sum_{m=1}^{M} z_m^k \sum_{i\in\mathcal{Y}} p_m^k(y=i) L_{mi} ||\boldsymbol{w_m}_j - \boldsymbol{v_m}_j^k||$ by using the property of the smooth function. Then, we have the divergence of each local step.

Lemma 2 (Divergence of each local step)**.** *We denote the distance between the local p_m^k and global p_m data distribution of modality m as $\triangle d_m^k = \sum_{i\in\mathcal{Y}} \parallel p_m^k(y=i) - p_m(y=i) \parallel$. Then, the divergence of each local step is bounded by:*

$$\sum_{m=1}^{M} z_m^k g_{max}\left(\boldsymbol{w_m}_{j-1}\right) \sum_{i\in\mathcal{Y}} B_{mi}^k \frac{\triangle d_m^k}{B_m^k} \left(\left(\eta B_m^k + 1\right)^{j-1-t_{p-1}} - 1\right) \tag{8}$$

where $g_{max}\left(\boldsymbol{w_m}_{j-1}\right) = \max_{i\in\mathcal{Y}} \parallel \nabla \psi_{(i)}\left(\boldsymbol{x_m}; \boldsymbol{w_m}_{j-1}\right) \parallel$ and $B_m^k = \sum_{i\in\mathcal{Y}} p_m^k(y=i) L_{mi} = \sum_{i\in\mathcal{Y}} B_{mi}^k$.

Proof: The weight divergence of each modality is derived separately. We first expand the $||\boldsymbol{w_m}_j - \boldsymbol{v_m}_j^k||$ based on the SGD update rule. Then, we add the zero term (i.e., $\nabla \psi_{(i)}(\boldsymbol{x_m}, \boldsymbol{w}_{(j-1)m})$ minus itself) and rearrange the expanded equation using the triangle inequality and the smoothness of the target function. The weight divergence in j-th iteration is computed by two terms, including the divergence of $j-1$ iteration and a gradient weighted by $\triangle d_m^k$. Finally, we utilize the Lemma 2 in work [24] to complete the mathematical induction for $j = [t_{p-1}, t_p]$ to obtain the upper bound. ■

Local-global divergence $\parallel \left(\nabla f_k\left(\boldsymbol{v}_j^k; \xi_j^k\right) - \nabla f\left(\boldsymbol{w}_j; \xi_j\right)\right) \parallel$ measures the distance between the local gradient and the global gradient.

Lemma 3 (Local-global gradient divergence)**.** *The local-global gradient divergence is bounded by:*

$$\sum_{m} z_m^k g_{max}\left(\boldsymbol{w_m}_j\right) \sum_{i\in\mathcal{Y}} \left(p_m^k(y=i) - p_m(y=i)\right) \tag{9}$$

where $g_{max}\left(\boldsymbol{w_m}_j\right) = \max_{i\in\mathcal{Y}} \parallel \nabla \psi_{(i)}\left(\boldsymbol{x_m}; \boldsymbol{w_m}_j\right) \parallel$.

Then, we can obtain the multi-modal weights divergence shown by Proposition 2.

Proposition 2 (Multi-modal weights divergence)**.** *After the p-th synchronization, the weights divergence between the aggregated global model $\boldsymbol{w}_{t_p}^{(f)}$ in multi-modal FL and the centralized trained model $\boldsymbol{w}_{t_p}^{(c)}$ follows the inequality below:*

$$\begin{aligned} \parallel \boldsymbol{w}_{t_p}^{(f)} - \boldsymbol{w}_{t_p}^{(c)} \parallel \leq & \parallel \boldsymbol{w}_{t_{p-1}}^{(f)} - \boldsymbol{w}_{t_{p-1}}^{(c)} \parallel \\ & + \sum_{k=1}^{K} p_k \sum_{j=1}^{E} (d_{local} + d_{local\_global}) \end{aligned} \tag{10}$$

where d_{local} is the upper bound of the $\parallel \nabla f_k(\boldsymbol{v}_j^k, \xi^k) - \nabla f_k(\boldsymbol{w}_j, \xi^k) \parallel$ shown in lemma 2 and d_{local_global} is the upper bound of $\parallel \nabla f_k(\boldsymbol{v}_j^k, \xi^k) - \nabla f(\boldsymbol{w}_j, \xi) \parallel$ shown in lemma 3.

C. Analysis

Proposition 2 shows that the global weight divergence of multi-modal FL is the collection of divergences in hierarchical training stages. Therefore, each level's over-fitting and inconsistent generalization rate is able to increase the global divergence.

As shown by lemma 2 and 3, in the local update, the weight divergence can be dominated by one modality m if its sub-network overfits on the local distribution that biases a lot from the global distribution, i.e. $\triangle d_m^k$. Different local distributions of M modalities can lead to inconsistent generalization rates among sub-networks, which leads to divergence accumulation, as shown by lemma 2. g_{max} further shows that the large gradient of either one sub-network exacerbates the weight divergence. Also, weight divergences of participating clients are accumulated. Therefore, a large weight divergence of any client can significantly increase global divergence.

As presented by Eq. (8), Eq. (9), and Eq. (10), the weight divergence of each modality and each client is re-weighted by the z_m^k and p_m, respectively. Therefore, these two parameters can be tuned to reduce the global divergence based on the overfitting and generalization behaviors.

V. Hierarchical Gradient Blending

With these insights, we propose a new algorithm, named hierarchical gradient blending (HGB), to address the discussed challenges of multi-modal FL.

Following the analysis of proposition 2, the main idea of our algorithm is to control the modality weights $\{z_m\}_{m=1}^{M}$ and aggregation weights $\{p_k\}_{k=1}^{K}$ to suppress the influence of overfitting on the weight divergence while balancing generalization rates in both the local update and global aggregation.

A. Theoretical Analysis of HGB

The learning target is to update the model to reduce the training loss while achieving low evaluation loss. Thus, HGB directly minimizes the overfitting-to-generalization ratio (OGR) shown in Eq. (4). In the multi-modal FL training process, our objective function is to obtain the best OGR for adjacent global weights $\boldsymbol{w}_{t_{p-1}}$ and $\boldsymbol{w}_{t_p}$ obtained by aggregating local models from K clients.

$$\min_{\{z_m\}_{m=1}^{M},\{p_k\}_{k=1}^{K}} \left(\frac{[L^T(\boldsymbol{w}_{t_{p-1}} - L^T(\boldsymbol{w}_{t_p})] - [L^*(\boldsymbol{w}_{t_{p-1}}) - L^*(\boldsymbol{w}_{t_p})]}{L^*(\boldsymbol{w}_{t_{p-1}}) - L^*(\boldsymbol{w}_{t_p})} \right)^2$$

Then, the approximate optimization problem of our objective function is presented in lemma 4.

Lemma 4. *For any global aggregation stage $t_p \in \mathcal{I}_e$, under the non-IID multi-modal data setting of K clients, the approximate optimization problem of our objective function is given as:*

$$\min_{\{z_m\}_{m=1}^{M},\{p_k\}_{k=1}^{K}} \frac{\sum_{k=1}^{K} p_k \sum_{j=t_{p-1}}^{t_p} \eta_j < \nabla L_k^T(\boldsymbol{v}_{\boldsymbol{m}_j}^k) - \nabla L_k^*(\boldsymbol{v}_{\boldsymbol{m}_j}^k), \sum_{m=1}^{M} z_m^k \boldsymbol{g}_{\boldsymbol{m}_j}^k >}{\sum_{k=1}^{K} p_k \sum_{j=t_{p-1}}^{t_p} \eta_j < \nabla L_k^*(\boldsymbol{v}_{\boldsymbol{m}_j}^k), \sum_{m=1}^{M} z_m^k \boldsymbol{g}_{\boldsymbol{m}_j}^k >}$$

where $j \in [t_{p-1}, t_p]$ and η_j is the learning rate of j-th step update. $\boldsymbol{g}_{\boldsymbol{m}_j}^k$ is the computed j-th step gradient of m modality sub-network in client k. L_k^T and L_k^* are the training loss and the "true" loss of the local distribution in client k, respectively.

Proof: The main tools used in our proof are Taylor theorem, recursion, and Jensen inequality. Firstly, we get the Taylor expansion of $L^T(\boldsymbol{v}_{\boldsymbol{m}_j})$ and $L^*(\boldsymbol{v}_{\boldsymbol{m}_j})$ where $j \in [t_{p-1}, t_p - 1]$ is the local iteration index. Then, with the recursion, we obtain the $L_k^T(\boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k)$ and $L_k^*(\boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k)$ which are then aggregated with p_k to get the $\sum_k p_k L_k^T(\boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k)$ and $\sum_k p_k L_k^*(\boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k)$. Secondly, using Jensen inequality, we get the relation $L^T(\sum_k p_k \boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k) \leq \sum_k p_k L_k^T(\boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k)$, which induces $L^T(\sum_k P_k \boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k) - L^T(\sum_k P_k \boldsymbol{v}_{\boldsymbol{m}_{t_{p-1}}}^k) \leq \sum_k P_k \sum_j \eta_j < \nabla L_k^T(\boldsymbol{v}_{\boldsymbol{m}_j}^k), \boldsymbol{g}_j^k >$. Likewise for the L^* terms. Finally, we can obtain the upper bound of the original objective function by putting the derived terms back to the equation. ■

The optimization problem in lemma 4 can be solved by the conclusion in theorem 1.

Theorem 1 (Optimal hierarchical gradient blending)**.** *According to the training procedure of the federated learning framework, the optimization problem in lemma 4 can be regarded as the combination of computing the optimal $\{z_m^k\}_{m=1}^{M}$, $k \in [1, K]$ in the local updates and computing the optimal $\{p_k\}_{k=1}^{K}$ in the global aggregation.*

In the local update, we can regard the $\{\boldsymbol{g}_{\boldsymbol{m}_j}^k\}_{m=1}^{M}$ as a set of estimates for L_k^* in the client k whose overfitting satisfies $E\left[\langle \nabla L_k^T - \nabla L_k^*, \boldsymbol{g}_{\boldsymbol{n}_j}^k \rangle \langle \nabla L_k^T - \nabla L_k^*, \boldsymbol{g}_{\boldsymbol{q}_j}^k \rangle\right] = 0$ if $n \neq q$ and $n, q \in M$. Also, in the global aggregation, the gradient of each client can be regard as one estimate for the L^* computed on the whole data. Besides, K estimates satisfies $E\left[\langle \nabla L_n^T - \nabla L_n^*, g_j^n \rangle \langle \nabla L_q^T - \nabla L_q^*, g_j^q \rangle\right] = 0$ if $n \neq q$ and $n, q \in K$.

The optimal $\{z_m^{k*}\}_{m=1}^{M}$ and $\{p_k^*\}_{k=1}^{K}$ are computed by:

$$z_m^{k*} = \frac{1}{Q} \frac{\langle \nabla l_k^*, \boldsymbol{g}_{\boldsymbol{m}}^k \rangle}{\sigma_m^2}, Q = \frac{\sum_{m=1}^{M} \frac{\langle \nabla l_k^*, \boldsymbol{g}_{\boldsymbol{m}}^k \rangle}{\sigma_m^2}}{2} \tag{13}$$

where $\sigma_m^2 = E[\langle \nabla l_k^T - \nabla l_k^*, \boldsymbol{g}_{\boldsymbol{m}}^k \rangle^2]$.

$$p_k^* = \frac{1}{M} \frac{\triangle G^k(t_{p-1}, t_p)}{2\left(\triangle O^k(t_{p-1}, t_p)\right)^2}, M = \sum_{k=1}^{K} \frac{\triangle G^k(t_{p-1}, t_p)}{2\left(\triangle O^k(t_{p-1}, t_p)\right)^2} \tag{14}$$

where $\triangle G^k$ and $\triangle O$ are the OGR terms of client k.

Finally, the z_m^* in the server is computed as:

$$z_m^* = \frac{\sum_{k=1}^{K} z_m^{k*}}{\sum_{m=1}^{M} \sum_{k=1}^{K} z_m^{k*}} \tag{15}$$

Proof: We omit the full proof due to space constraints. For both the optimization problem in the local update and the global aggregation, we use $(\sum_n a_n)^2 = \sum_n a_n^2 + \sum_i \sum_{j \neq i} a_i a_j$ and the assumption to remove the cross terms. Also, without loss of generality, we use the constrains $\sum_{m=1}^{M} z_m^k = 1$ and $\sum_{k=1}^{K} p_k = 1$. Thus, we can apply lagrange multipliers to the corresponding objective function, which can then be solved directly by setting its gradient to zero. Then, Eq. (15) is obtained directly by calculating votes on each modality m from K clients. ■

Finally, we can minimize the OGR of our multi-modal FL by using the optimal weights shown in Eq. (13) and Eq. (14). And, our mathematical analysis presents that the parameters of HGB can be adaptively computed according to the overfitting and generalization conditions during the learning process. Besides, through Eq. (15), we can conclude that modality sub-network with high generalization is desired to be assigned higher z_m^*.

B. HGB in Practice

In practice, one basic setting is that the l^* can be approximated by the loss l^V computed on the validation set, which agrees with work [21]. Thus, l_k^* of each client k is measured based on its validation set D'^k. Besides, we utilize the cross-entropy loss function as discussed in proposition 2.

$\{z_m^{k*}\}_{m=1}^{M}$ is obtained at the end of E local steps in client k. Based on Eq. (4), Eq. (5) and our discussion in the theorem 1, σ_m can be replaced by the overfitting term $\triangle O_m^k(t_{p-1}, t_p)$ of modality m between the initialization weight $\boldsymbol{v}_{\boldsymbol{m}_{t_{p-1}}}^k$ and the updated weights $\boldsymbol{v}_{\boldsymbol{m}_{t_p}}^k$. Also, the $\langle \nabla l_k^*, \boldsymbol{g}_{\boldsymbol{m}}^k \rangle$ is replaced by the generalization term $\triangle \boldsymbol{g}_{\boldsymbol{m}}^k(t_{p-1}, t_p)$ of the modality m. The losses used for computing $\triangle O_m^k$ and $\triangle \boldsymbol{g}_{\boldsymbol{m}}^k$ are obtained by

applying the corresponding weight $\boldsymbol{v_m^k}$ to the trainset D^k and validation set D'^k. Thus, in the local update stage, K clients can obtain their own $\left\{z_m^{k*}\right\}_{m=1}^M$ in parallel.

For $\{p_k^*\}_{k=1}^K$, to compute Eq. (14), each client requires to report its own overfitting rate $\triangle O^k$ and generalization rate $\triangle G^k$ to the server. For each client k, we implement the optimal gradient blending by loss re-weighting formatted as $\left(L_k^T\right)_{blend} = \sum_{m=1}^M z_m l_{km}^T$ where l_{km}^T is the training loss of the modality m. Likewise for the validation loss $\left(L_k^V\right)_{blend}$. Finally, these blended losses are used to measure $\triangle O^k(t_{p-1}, t_p)$ and $\triangle G^k(t_{p-1}, t_p)$.

Applying weights to the whole training set or validation set can be expensive. Thus, we can compute losses based on the subsets D_s^k and $D_s'^k$.

The detailed procedure of HGB in the multi-modal FL is shown in algorithm 1.

C. Convergence Analysis

As described in the algorithm 1, the HGB computes the optimal weights for the modality sub-networks and local models online. According to our theorem 1, the only requirement of the online computation is to obtain the training and validation errors of the updated model. This prevents HGB from introducing additional trainable parameters into the objective function and training architecture. Therefore, our HGB is both task-agnostic and architecture-agnostic.

Following our discussion Proposition 1, non-IID multi-modal data among clients leads to the unbounded gradient. Also, the bound of gradient variance shown in lemma 1 induces that the convergence depends on the meaningful quantity $\sum_{k=1}^K p_k E \parallel \nabla f_k\left(\boldsymbol{v}_*^k; \xi\right) \parallel^2$. Under the convex and smooth assumptions, the general convergence statement has been made in Corollary 1 of the work [25]. The convergence rate $\frac{1}{\sqrt{KT}}$ of HGB in multi-modal FL can be guaranteed by using $E = O\left(T^{\frac{1}{4}} K^{-\frac{3}{4}}\right)$ where K is the number of selected clients in each round.

Algorithm 1: Multi-modal FL with Hierarchical Gradient Blending

Input: initial Learning rate η_0, local update step E and batch size B, total round P, total client C with index k, subset percent S, number of modalities M.

Output: $\{\boldsymbol{w_m}_P, z_{mP}^*\}_{m=1}^M$.

Initialization: $t_0 \leftarrow 0$, initialize $\{\boldsymbol{w_m}_0\}_{m=1}^M$, $z_{m0} = \frac{1}{M}$.

- Main loop

for *each round* $p = 1, 2, ..., P$ **do**

  $C_{t_p} \leftarrow$ random set of K clients from total C clients

  Distribute $\left\{\boldsymbol{w_m}_{t_{p-1}}, z_{mt_{p-1}}^*\right\}_{m=1}^M$ to C_{t_p}

  for *each client* $k \in C_t$ *in parallel* **do**

    Operate clientOperation.

    Send $\left\{\boldsymbol{w_m}_{t_p}^k, z_{mt_p}^k\right\}_{m=1}^M, \triangle O^k, \triangle G^k$ to server.

  end

  Compute p_k^* using $\left\{\triangle O^k\right\}_{k=1}^K$ and $\left\{\triangle G^k\right\}_{k=1}^K$ according to Eq. (14);

  Compute $\boldsymbol{w_m}_{t_p} = \sum_{k=1}^P p_k^* \boldsymbol{w_m}_{t_p}^k$ and $z_{mt_p}^*$.

end

Function clientOperation $(\{\boldsymbol{w}_m, z_m^*\}_{m=1}^M)$ **:**

  $j \leftarrow 0$, $\boldsymbol{v_m}_j = w_m$

  $D_s, D_s' \leftarrow$ Random S percent subset from D, D'.

  For each modality m, compute $l_m^T\left(\boldsymbol{v_m}_j\right)$, $l_m^V\left(\boldsymbol{v_m}_j\right)$.

  for *each step* $j = 1, 2, ..., E$ **do**

    $\boldsymbol{b} \leftarrow$ sample a batch of data from D.

    $\boldsymbol{v_m}_{j+1} = \boldsymbol{v_m}_j - \eta_j \nabla \left(\sum_{m=1}^M z_m l_m\left(\boldsymbol{v_m}_j\right)\right)$.

  end

  For each modality m, compute $l_m^T\left(\boldsymbol{v_m}_E\right)$, $l_m^V\left(\boldsymbol{v_m}_E\right)$

  Compute $z_m^* = \frac{1}{Q} \frac{\triangle G_m(0,E)}{(\triangle O_m(0,E))^2}$, where $Q = \sum_{m=1}^M \frac{\triangle G_m(0,E)}{2(\triangle O_m(0,E))^2}$.

  Compute $\triangle G(0,E)$ and $\triangle O(0,E)$ with $l^V = \sum_{m=1}^M z_m^* l_m^V$.

  return $\left\{\boldsymbol{v_m}_E^k, z_m^*\right\}_{m=1}^M, \triangle O, \triangle G$

VI. Experimental Results

In this section, we evaluate the performance of the proposed HGB training algorithm in multi-modal FL. Based on a total of $M = 3$ modalities, including RGB frames (RGB), optical flow (OF), and audio (A), the multi-modal model is trained to complete the video classification task. The test platform used in all experiments is Plato, a software framework to facilitate scalable federated learning research.

A. Experimental Setup

Datasets. We utilize Kinetics [26] and Finegym [27] datasets. Kinetics is a standard benchmark for action recognition with $260k$ videos of 400 human action classes. There are $240k$ and $20k$ videos in the train split and validation split, respectively. For the Finegym dataset, we utilize the Gym99 containing 20484 and 8521 element-level action instances in the train split and validation split, respectively.

Learning Setting. Our experiments focus on the video classification task. There are a total of $C = 50$ clients. The maximum number of communication rounds is $P = 1000$. Then, in each round, $K = 40$ clients are randomly selected to participate in the training, and each client runs $E = 300$ steps mini-batch SGD in the local update with a momentum of 0.9. The corresponding batch size is 24. The learning rate of each client is 0.000125, while the weight decay is 0.001. The 40% percent of the dataset is pulled from the local train set and local validation set to obtain L^T and L^V.

Model. We use Recognizer3D with 50 layers as the visual backbone for RGB and optical flow (OF). Then the AudioRecognizer with 50 layers is used to process the audio (A) data. The

classification heads of all modalities share the same structure- a two-layer fully connected network with hidden dimension 512. Then, the modality fusion part is designed as a network with two fully-connected layers. The features from visual and audio backbones are concatenated to feed into the prediction layer with dimension 512. Therefore, we have four multi-modal models that process different modality combinations, including A+RGB, OF+RGB, A+OF, A+OF+RGB. For simplicity, we use these abbreviations to represent the corresponding multi-modal models. Also, the corresponding uni-modal models include A, RGB, and OF. For RGB and flow, we use clips of $16\times224\times224$ as input. We follow CSN [28] for visual pre-processing and augmentation. As for the audio, we use log-Mel with 100 temporal frames by 40 Mel filters. Audio and visuals are temporally aligned.

Benchmark methods. The FL methods, including the FedAvg [1], FedAttn [16], and FedNova [17], are used as the benchmark to compare with HGB. The use of HGB in the FL paradigm is referred to as **FedHGB** in our experiments. Besides, we use the corresponding best uni-modal model model as the baseline.

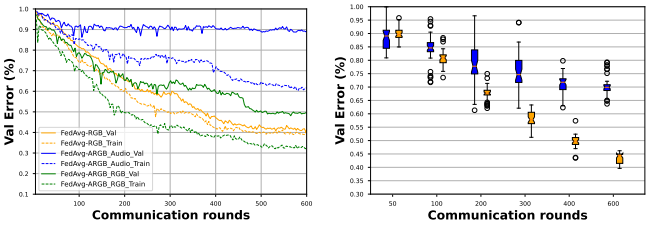
Non-IID multi-modal data. We consider three different ways (Case A, B, C) of distributing the data to clients, thereby simulating the non-IID multi-modal data among participating clients. For all the cases, clients contain all classes but with different distributions. Thus, the basic setting is the distribution-based label non-IID implemented by Dirichlet distribution with concentration parameter 0.5. In case A, three modalities are uniformly assigned to each client, making each sample contains all modalities. In case B, we achieve the quantity-based modality non-IID in which each client can only contain subset modalities. It has three types, including $\text{mixed}-B$, $2M-B$, and $1M-B$. In $\text{mixed}-B$, the number of modalities in each client is in the range $[1, 3]$ while $2M-B$ and $1M-B$ contain 2 and 1 modalities, respectively. Case C is built based on case $\text{mixed}-B$. We additionally add the sample skewness among modalities. Therefore, datasets of different modalities have different sizes, leading to modality incompleteness.

Performance metrics: The video classification top-1 accuracy ($V@1$) and the number of communication rounds (CR) required to convergence are used to present the performance of the training algorithm.

B. Inconsistent Overfitting and Generalization Rates

We first apply the FedAvg method to the uni-modal FL and the multi-modal FL to present the challenges of training the multi-modal model in the non-IID multi-modal data. Using the Kinetics dataset with non-IID case $1M$-B, we compare the audio-RGB (ARGB) model with the uni-modal RGB-only model, i.e., FedAvg-RGB and FedAvg-ARGB, respectively.

Fig. 1 (a) plots the training curve and the validation curve on Kinetics. The over-fitting problem of FedAvg-ARGB containing two sub-networks is far more severe than that of FedAvg-RGB. Specifically, training the audio-RGB model achieves lower training error and higher validation error than the RGB-only model, inducing the accuracy drop shown in Table I. The



(a) (b)

Fig. 1: **Severe overfitting and inconsistent generalization rate of audio-RGB multi-model on Kinetics with non-IID case B.** Solid lines plot validation error while dashed lines show train error.

audio network and RGB network of FedAvg-ARGB show significantly different generalization rates. The RGB network generalizes better than that of the audio network. Therefore, the low convergence speed and accuracy in Table I can be further explained by joint training two sub-networks that generalize at different rates.

Fig. 1 (b) shows the generalization errors of participating clients in the training process. The generalization variance among local models of FedAvg-ARGB maintains a particularly higher value than the FedAvg-RGB. Especially in communication round 300, the variance of generalization errors is ten times higher than that of FedAvg-RGB. Thus, the inconsistent generalization rates among local models lead to an unstable learning process and a sub-optimal model. This explains the significant performance degradation for multi-modal FL in Table I.

TABLE I: The performance comparison of the FedAvg method on uni-modal FL and the multi-modal FL under the non-IID data settings of three modalities.

	Centralized	FedAvg	
Modalities	V@1	V@1	#Rounds
RGB	71.52	58.43	480
A+RGB	72.17	49.91	519
OF+RGB	72.3	52.57	504
A+OF+RGB	73.62	38.62	557

C. Comparison with State-of-the-Art

This section shows the quantitative results and analysis of HGB compared with state-of-the-art benchmark methods under non-IID cases A, B, and C. The performance of the listed methods is evaluated by applying them to train a multi-modal model with three modalities A+OF+RGB sub-networks.

Fig. 2 shows the validation curve in the training process, and Table II shows the validation accuracy in two non-IID types of case B. As shown by Fig. 2, the smooth validation curve demonstrates that our FedHGB can maintain stability in learning under all non-IID cases. In summary, on both the datasets with three non-IID cases A, B, and C, the performance

TABLE II: The performance comparison of methods in case B with two modality non-IID types (i.e., Mixed-B and 2M-B). The evaluation metric is the top-1 accuracy and the communication rounds distance ($\triangle CR$) between FedHGB and the fastest method.

Datasets	Kinetics		Gym	
Case B settings	Mixed-B	2M-B	Mixed-B	2M-B
FedAvg	38.04	44.32	42.33	51.42
FedAttn	51.79	56.91	58.07	64.52
FedNova	55.12	58.76	63.92	68.3
FedHGB	62.97	64.39	71.66	73.34
$\triangle CR$	34	15	51	20
Uni-RGB	62.33		70.52	

of our FedHGB outperforms other leading methods in terms of validation accuracy and convergency speed. In case A shown by Fig. 2 (a)(b), with only around 300 communication rounds (CRs), FedHGB achieves about 68% and 75% accuracy on the two datasets, outperforming other methods. In the more challenging case B in Table II, FedHGB utilizes minimum CRs to achieve average 6.745% and 6.44% improvements on two datasets over the best heterogeneous federated optimization method FedNova. Besides. In the case C shown by Fig. 2 (e)(f), the accuracy of our method is more than 10% higher than FedNova, yet with at least 100 CRs reduction compared with others. Also, shown by the last row of Table II, the performance of multi-modal trained by FedHGB consistently outperform the best uni-RGB model.

The experimental results that HGB is far superior to other methods demonstrate that in multi-modal FL, training the optimal multi-modal model with multiple sub-networks requires the optimal blending of modalities and reweighting of the local models. FedAvg that sole averages the local models with equal weights obtain the worst accuracy in all cases. Then, compared with FedAttn that reweights the local models based on the weight divergence, our FedHGB computes the optimal weights according to clients' generalization behaviors directly, leading to at least 15% higher validation accuracy. One main reason for this is presented by Eq. 10 in proposition 2. The discussion in FedAttn ignores that the weight divergence is also derived from the M sub-networks in multi-modal FL. The FedNova that aggregates the normalized stochastic gradients from local updates partly addresses the challenge in multi-modal FL because it suppresses the gradient divergence among clients. However, our FedHGB further induces the optimal blending of modalities in the local update, making it significantly outperform FedNova in cases B and C that mainly contain modality-based non-IID.

D. Ablation Experiments

This section illustrates the performance impact of optimal gradient blending in each layer of HGB. Specifically, we consider two ablations, including the M-GB and C-GB. The M-GB only computes the optimal blending of modalities z_m^* in the local update but fixes the p_k^* as the $1/K$. C-GB only computes p_k^* but fixes z_m^* as $1/M$. Our proposed theorem 1 supports these ablation experiments as the two blending levels can be computed separately to minimize the global OGR.

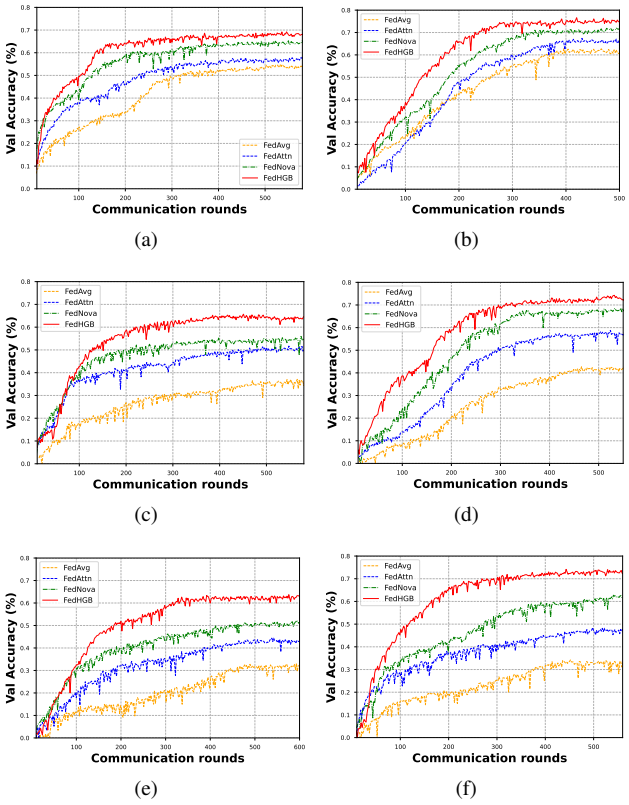


Fig. 2: **The validation curves on Kinetics *(i.e., (a), (c), (e))* and Gym *(i.e., (b), (d), (f))* datasets with three non-IID cases A, $1M$-B, and C.** FedHGB far surpasses other methods in non-IID case A, B, and C whose sub-figures are presented in the first row, second row, and third row, respectively.

As presented in Table III, compared with the full HGB, the validation accuracy obtained by solely applying M-GB or C-GB shows a significant performance decrease in two datasets under all cases. This demonstrates that simultaneously achieving the optimal blending of modalities and computing the aggregation weights of local models is the key point to efficiently training an effective multi-modal model under the non-IID multi-modal data. This verifies our theorem 1.

Then, the importance of optimal blending of modalities in multi-modal FL is shown by the competitive results of M-GB in all settings. For instance, under non-IID cases A and C, M-GB shows averaged 3.56% and 5.46% accuracy drop in two datasets, respectively. M-GB still maintains strong performance in challenging case B. The main reason is that in all cases of multi-modal FL, adjusting the gradient of different modality sub-networks is an inherent requirement for joint training. As each client contains a subset of modalities, there is a high degree of inconsistency among clients. The limited accuracy drop obtained by C-GB shows the necessity of reweighting

TABLE III: The performance comparison between ablation methods of HGB in two datasets with all non-IID settings. The evaluation metric is the top-1 accuracy (%) and the communication rounds $\triangle CR$ that is computed as CR of M-GB minus the CR of C-GB .

Datasets	methods	CaseA	mixed-B	2M-B	1M-B	Case C
Kinetics	M-GB	65.92	56.55	60.47	58.73	57.66
	C-GB	63.83	55.91	57.02	58.64	52.81
	$\triangle CR$	25	48	66	95	67
Gym	M-GB	71.36	66.13	69.92	67.34	65.17
	C-GB	70.03	64.4	66.1	66.38	58.93
	$\triangle CR$	41	36	69	87	46

local models to alleviate the gradient divergence. However, M-GB still obtains accuracy close to C-GB. We argue that M-GB can suppress gradients of sub-networks that damage local generalization, making the weight update direction of each client contribute to global generalization.

Convergence speed comparisons of case B in Table III demonstrate that C-GB with local updates reweighting can reduce communication rounds by 70 on average compared to M-GB while maintaining a lower accuracy drop of 1.39%. However, in case C with the modality incompleteness, C-GB presents a 4.85% accuracy drop compared with M-GB. We believe that training in case C requires the blending of modalities. However, unlike M-GB, C-GB does not compute the optimal blending of gradients from sub-networks in each local update step to reduce generalization error, leading to its low accuracy.

E. Qualitative Analysis

Our analysis of qualitative results in this section aims to present the effectiveness of the modality blending in the local update and the relation between the generalization rate and the client's weight p_k.

As shown by the first column of Fig. 3, with the optimal blending of modalities in the local update, gradients that damage the generalization are alleviated, making the local model is updated toward the minimum OGR direction as described in our theorem 1. Thus, the generalization error variance of updated local models in FedHGB is significantly lower than the variance in FedAttn.

The relationship between the generalization error of the client and the assigned weight, shown in Fig. 3 demonstrates that FedHGB tends to assign higher weights to those clients with low generalization errors, especially when $CR \in [150, 200]$. This contributes to training a global model that behaves well on the whole population. On the contrary, the weights computed by FedAttn are independent of the generalization condition, leading to its low validation accuracy. Thus, we can empirically conclude that the best performance of FedHGB is derived from consistently focusing on tuning weights for low generalization error.

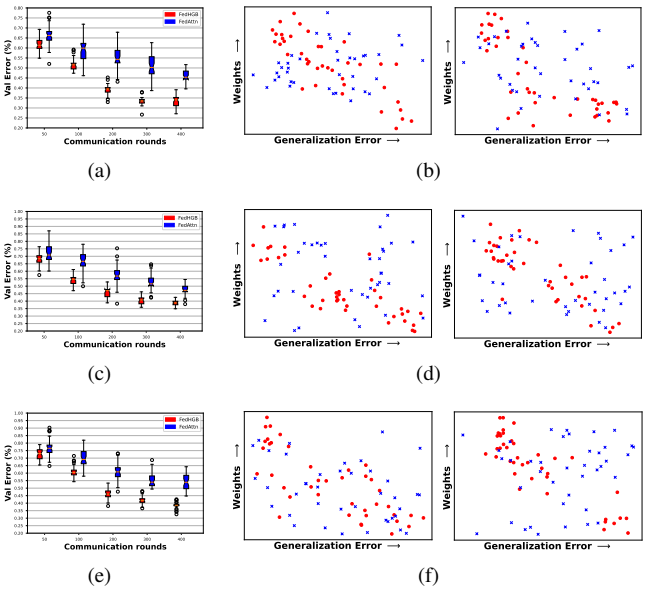


Fig. 3: **Comparison of quantitative results between FedHGB and FedAttn on the Kinetics dataset, where the three rows represent the three non-IID cases A, $\text{mixed}-B$, and C, respectively.** The first column shows the generalization distribution of clients before aggregation in different communication rounds. The first column shows the generalization distribution of clients before aggregation in different communication rounds. The second column in the right subfigure shows the relationship between generalization error and the computed weight p_k for participating clients in the round $50-100$. The third column shows the corresponding relation in the round $150-200$.

VII. Concluding Remarks

In this paper, we investigated the challenges involved as multi-modal models are trained in federated learning (FL) with non-IID multi-modal data. The primary challenge we focused on is that both the local updates and global aggregation suffer from overfitting and inconsistent generalization rates, which causes significant degradation in the performance of existing FL methods. A highlight in our original contributions is a new training algorithm, referred to as hierarchical gradient blending (HGB), which adaptively achieves the optimal blending of modality sub-networks and the optimal aggregation of local updates. Notably, HGB's design seeks to minimize the overfitting-to-generalization rate (OGR) of the global model. We have presented a rigorous theoretical analysis to prove that HGB guarantees effectiveness in the context of multi-modal FL. Based on such a design, the blending weights are computed online based on the overfitting and generalization behaviors. Our extensive experimental results on video classification datasets validated our theoretical analysis and demonstrated the effectiveness of HGB in a variety of non-IID multi-modal data scenarios.

Acknowledgment Sijia Chen was supported by NSERC Discovery Research Program.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[4] S. Caldas, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

[5] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[6] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "Fedvision: An online visual object detection platform powered by federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 08, 2020, pp. 13 172–13 179.

[7] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[8] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1698–1707.

[9] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2018.

[11] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Federated learning for vision-and-language grounding problems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 572–11 579.

[12] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[13] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," 2018.

[14] M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar, "Adaptive gradient-based meta-learning methods," in *Advances in Neural Information Processing Systems*, 2019, pp. 5917–5928.

[15] T. Yu, E. Bagdasaryan, and V. Shmatikov, "Salvaging federated learning by local adaptation," 2020.

[16] S. Ji, S. Pan, G. Long, X. Li, J. Jiang, and Z. Huang, "Learning private neural language modeling with attentive aggregation," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[17] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *arXiv preprint arXiv:2007.07481*, 2020.

[18] X. Yao, T. Huang, R.-X. Zhang, R. Li, and L. Sun, "Federated learning with unbiased gradient aggregation and controllable meta updating," *arXiv preprint arXiv:1910.08234*, 2019.

[19] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," *arXiv preprint arXiv:2002.05516*, 2020.

[20] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," *arXiv preprint arXiv:1902.00146*, 2019.

[21] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 695–12 705.

[22] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 794–803.

[23] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[24] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 63–71.

[25] A. Khaled, K. Mishchenko, and P. Richtárik, "First analysis of local gd on heterogeneous data," *arXiv preprint arXiv:1909.04715*, 2019.

[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[27] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2616–2625.

[28] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.