# *Supplementary Material* for Towards Multi-modal Transformers in Federated Learning

Guangyu Sun[1], Matias Mendieta[1], Aritra Dutta[2], Xin Li[2], and Chen Chen[1]

[1] Center for Research in Computer Vision, University of Central Florida, FL, USA
[2] Department of Mathematics, University of Central Florida, FL, USA
{guangyu, matias.mendieta, aritra.dutta, xin.li}@ucf.edu;
chen.chen@crcv.ucf.edu

The supplementary material is structured as follows:

1. In Section A we sketch a general framework for transfer MFL (§A.1), then we adopt it to the transfer MFL setting in the vision-language (§A.2). We give the convergence guarantee in §A.3. Finally, in §A.4, we argue why multi-modal FL might work with diverse data modalities by providing a simple heuristic on the distribution of the multimodal data.
2. Section B elaborates on more experimental details, including (*i*) implementation details, (*ii*) visualization of the data partitioning, (*iii*) stochasticity discussion, (*iv*) breakdown performance of the results in the main paper, (*v*) communication analysis, (*vi*) experiments under imbalanced total client numbers, and (*vii*) visualizations.
3. Section C discusses the potential negative social impact and limitation of the proposed method.

## A Theoretical Guarantee

In this section, we start with a general transfer multimodal FL framework.

### A.1 A general framework for transfer MFL

In this section, we outline the general algorithmic framework of transfer MFL. Consider the *empirical risk minimization* (ERM) problem:

$$\min_{w\in\mathbb{R}^d}\left[\mathcal{F}(w) := \sum_{j=1}^{M} F_j(w) = \sum_{j=1}^{M}\left(\underbrace{\frac{1}{N_j}\sum_{i=1}^{N_j} f_{ji}(w)}_{:=F_j}\right)\right], \tag{1}$$

where $f_{ji}(w) = \frac{1}{|D_j|}\sum_{i\in D_j}\mathbb{E}_{z\sim\mathcal{D}_i} l(w;z)$ denotes the loss function evaluated on input datapoint at the $i^{\text{th}}$ client, with $D_j$ denoting the set of all clients whose data has non-empty $j$th modality, $M$ is the total number of modality types, and $N_j$ denotes the number of clients for the $j^{\text{th}}$ modality. At the beginning of the

training process, the server initializes global models, $w^{(0)}$. The entire process runs for $T$ global communication rounds. At the beginning of each global round, $r$, each client, $i$, receives the global model, initializes it to $w_i^0 = w^{(r)}$, and updates its local model in each iteration, $t$. We compactly write it:

$$w_i^{t+1} = w_i^t - \Omega_{\text{local}} \nabla w_i^t, \tag{2}$$

where $\Omega_{\text{local}}$ is a block diagonal matrix, where each block is of the size of the parameters of that individual modality type. Each client completes local training (2), for $E$ local epochs, communicates the model update to the server, and the server aggregates them at the $r^{\text{th}}$ round via:

$$\nabla w^{(r)} = \Omega_{\text{Aggregation}}(w_i^E - w^{(r)}). \tag{3}$$

Finally, the update rule at the server for each round $r$ is given as

$$w^{(r+1)} = w^{(r)} + \Omega_{\text{server}} \nabla w^{(r)}, \tag{4}$$

where $\Omega_{\text{server}}$ is a symmetric, $(M+1) \times (M+1)$ block band matrix.

### A.2 Transfer MFL in our setting

We consider a transfer multi-modal FL setting in the vision-language domain with a total of $N$ clients. Based on the general framework, we describe our problem setup. Although (1) presents a general multimodal loss function, in our case, we are working with 3 modality types, so, $M = 3$. Let $D_{\text{v}}, D_{\text{l}}$, and $D_{\text{vl}}$ present the dataset for vision, text, and multimodal tasks, respectively. Note that, $D_{\text{vl}} = \{d : d = (v, l)\}$ contains vision and text pair as input, and define $D_{\text{vl}}(v) := P_v(D_{\text{vl}})$, $D_{\text{vl}}(l) := P_l(D_{\text{vl}})$, where $P_y$ is the projection operator on the set $y$. For training the vision models, we use $D_{\text{v}} \cup D_{\text{vl}}(v)$; for training the language models, we use $D_{\text{l}} \cup D_{\text{vl}}(l)$. At the beginning of the FL process, the server initializes global models, $w^{(0)} := \left(w^{(0,\text{v})}, w^{(0,\text{l})}, w^{(0,\text{vl}_\text{v})}, w^{(0,\text{vl}_\text{l})}\right)^\top$ the set of parameters, and $\nabla w^r = \left(\nabla w_v^r \; \nabla w_l^r \; \nabla w_{vl_v}^r \; \nabla w_{vl_l}^r\right)^\top$.

For local training, $\Omega_{\text{local}} = \text{diag}(\eta_{\text{v}} \text{I}_{|\text{D}_\text{v}|} \; \eta_{\text{l}} \text{I}_{|\text{D}_\text{l}|} \; \eta_{\text{vl}} \text{I}_{|\text{D}_\text{vl}(\text{v})|} \; \eta_{\text{vl}} \text{I}_{|\text{D}_\text{vl}(\text{l})|})$, where $\eta_v, \eta_l, \eta_{vl} \geq 0$ are stepsizes for respective modality type. Note that, $|x|$ denotes the dimension of an arbitrary vector, $x$, and $I_{|x|}$ is the identity matrix in the space $\mathbb{R}^{|x| \times |x|}$. Let $n_v, n_l$ and $n_{vl}$ clients be sampled uniformly at random from each $\{N_j\}_{j=1}^3$, respectively, in each training round. Hence, after $E$ local epochs, (3) is given as:

$$\begin{cases} \nabla w_v^r := \frac{1}{n_v} \sum_{i \in n_v} (w_{v_i}^E - w_v^r), \nabla w_l^r := \frac{1}{n_l} \sum_{i \in n_l} (w_{l_i}^E - w_l^r), \\ \begin{pmatrix} \nabla w_{vl_v}^r \\ \nabla w_{vl_l}^r \end{pmatrix} := \frac{1}{n_{vl}} \sum_{i \in n_{vl}} \begin{pmatrix} w_{vl_{v\,i}}^E - w_{vl_v}^r \\ w_{vl_{l\,i}}^E - w_{vl_l}^r \end{pmatrix}. \end{cases}$$

Our setting allows different structures of $\Omega_{\text{server}}$ that can be adapted to consider cross-modal contribution. Specifically, we consider $\Omega_{\text{server}}$ as follows: For

$i$ odd, $(\Omega_{\text{server}})_{ii} = A_v$, and for $i$ even, $(\Omega_{\text{server}})_{ii} = B_l$. The superdiagonal blocks, $j > i$, are given as: $(i)$ For $i$ odd and $j$ even, $(\Omega_{\text{server}})_{ij} = 0$; for $j$ odd, $(\Omega_{\text{server}})_{ij} = B_{vl}$; $(ii)$ For $i$ even and $j$ odd, $(\Omega_{\text{server}})_{ii} = 0$, for $j$ even, $(\Omega_{\text{server}})_{ij} = D_{\text{vl}}$. Constructing $\Omega_{\text{server}}$ as above allows us to leverage the participation and interaction of each modality over the *vanilla* multimodal FedAvg. For *vanilla* FedAvg [4, 7], $\Omega_{\text{server}}$ is a block diagonal matrix and $n_v = n_l = n_{vl}$.

### A.3 Convergence guarantee

Based on the convergence of FedAvg in [4, 7], in this section, we will comment on the convergence of general transfer MFL. For ease of notation we consider, at each round, $S_j$ be the number of clients sampled for the the $j^{\text{th}}$ modality.

**Assumptions.** We require the following assumptions.

**Assumption 1** ***(Global minimum)*** *For each $j \in [M]$, there exists $w^\star$ such that, $F_j(w^\star) = F_j^\star \leq F_j(w)$, for all $w \in \mathbb{R}^d$.*

**Assumption 2** ***($\beta$-Smoothness)*** *The loss function $f_{ij} : \mathbb{R}^d \to \mathbb{R}$ at each node is $\beta$-smooth, i.e. $f_{ij}(y) \leq f_{ij}(x) + \nabla f_{ij}(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^d$.*

*Remark 1.* The above assumption implies that $F_j$ is $\beta$-smooth for all $j$.

**Assumption 3** *For each $j \in [M]$, there exist constants $G_j \geq 0, B_j \geq 1$, such that for all $x \in \mathbb{R}^d$, the stochastic noise, $\xi_{i,t}$ follows*

$$\frac{1}{N_j} \sum_{i=1}^{N_j} \|\nabla f_{ij}(x)\|^2 \leq G_j^2 + B_j^2 \|F_j(x)\|^2.$$

**Assumption 4** ***(Bounded variance)*** *For each $j \in [M]$, let $g_{ji}(w) := \nabla f_{ji}(w, z_{i(k)})$ be the unbiased stochastic gradient of $f_{ji}$ with bounded variance. That is, there exists, $\sigma_j \geq 0$ such that, $\mathbb{E}_{z_{i(k)}} \left[\|g_{ji}(w) - \nabla f_{ji}(w)\|^2\right] \leq \sigma_j^2$, for all $w, i$, where $z_{i(k)}$ is the $k^{\text{th}}$ sample data at the $i^{\text{th}}$ node.*

**Assumption 5** *The eigenvalues of the symmetric matrix, $\Omega_{\text{server}}$ are nonnegative.*

*Remark 2.* The above assumption implies that there exists an orthogonal matrix $P$ such that, $\Omega_{\text{server}} = P \Lambda P^\top$, where $\Lambda$ is a diagonal matrix of nonnegative eigenvalues of $\Omega_{\text{server}}$.

*Remark 3.* Based on the previous remark, the update rule (4) can be rewritten as:

$$P^\top w^{(r+1)} = P^\top w^{(r)} + P^\top P \Lambda P^\top \nabla w^{(r)}.$$

Consider the change of variable, $\tilde{w}^{(r)} := P^\top w^{(r)}$ and hence the above becomes:

$$\tilde{w}^{(r+1)} = \tilde{w}^{(r)} + \Lambda \nabla \tilde{w}^{(r)}. \tag{5}$$

Finally, we are all set to give our main convergence result based on the vanilla FedAvg framework; for more details see [4, 7].

**Theorem 1.** *For each $j \in [M]$, let $F_j$ satisfies Assumptions 1-5. Then*

$$\mathbb{E}\left[\|\nabla F_j(\tilde{w}^{(T)})\|^2\right] \leq O\left(\frac{\beta\sqrt{(F_j(\tilde{w}^{(0)}) - F_j^\star)}}{\sqrt{TES_j}}\right).$$

*Remark 4.* From (1), we have $\mathcal{F}(\tilde{w}) = \sum_{j=1}^{M} F_j(\tilde{w})$. Hence the boundedness of each $\mathbb{E}\left[\|\nabla F_j(\tilde{w}^{(T)})\|\right]$ guarantee the boundedness of $\mathbb{E}\left[\|\nabla \mathcal{F}(\tilde{w}^{(T)})\|\right]$.

### A.4 Distribution of the data

In this subsection, we discuss the perspective of multi-modal learning from the learning of the joint distribution of the modalities. In general, we could (and should) not assume independence among the different modalities at hand, and thus, their joint distribution is not simply the product of the marginal distributions. The training datasets then should be samples that reflect the same distributions of each modality feature. Let $\mathcal{D}$ be the joint *unknown* distribution of the input data of two modalities. Let the datasets, $D_{\mathrm{v}}$ and $D_{\mathrm{l}}$ have $\mathcal{D}_{\mathrm{v}}$ and $\mathcal{D}_{\mathrm{l}}$ as their marginal probability distributions of modalities $v$ and $l$ respectively. The availability of the dataset $D_{\mathrm{vl}}$ that follows the distribution $\mathcal{D}$ makes it possible to learn the joint distribution when $v$ and $l$ modalities are not independent; see Figure 1. The learning of a joint density model over the space of multimodal inputs is likely to yield a better generalization in various applications [10]. Intuitively, this explains the possibility that multi-modal learning could improve performance when modalities are jointly used in training the parameters even for the individual modality model.

## B Numerical Experiments

This section serves as an addendum to the numerical experiments in the original paper.

### B.1 Implementation Details

**General setup.** We use AdamW [6] as the optimizer with a learning rate of 0.0001 with a decay of 0.99 every epoch for local training. The batch size is set as 112 in most cases. All the experiments are implemented under the PyTorch framework and run on 4× Nvidia A5000 GPUs.

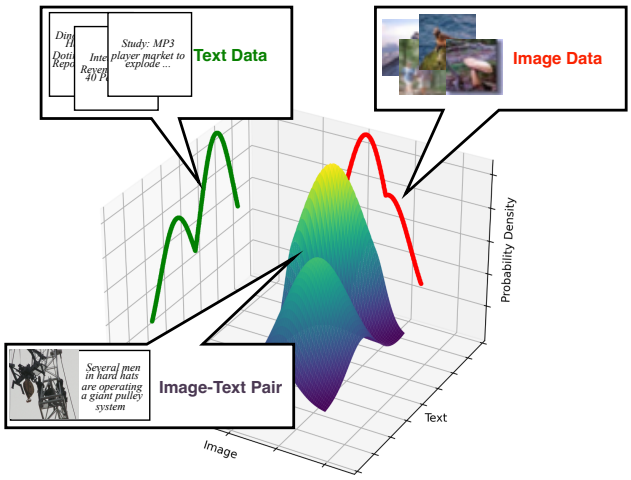



**Fig. 1:** Multi-modal FL in the vision-language domain with collaboration from different modalities.

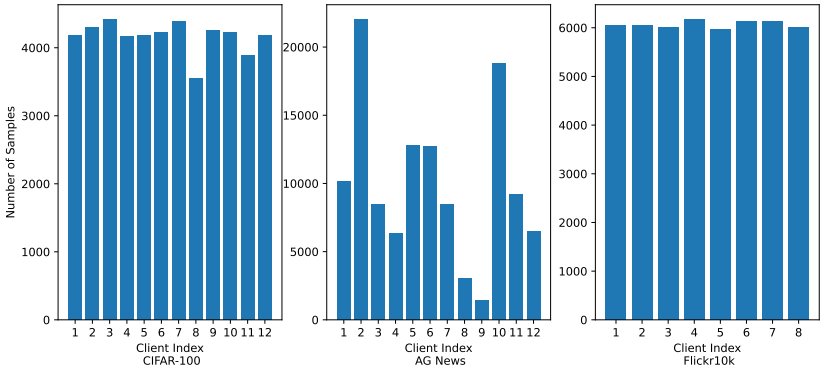



**Fig. 2:** Visualization of the data partitioning of different datasets: CIFAR-100 [5], AG News [13], Flickr10k [9].

**CreamFL [11].** In the original CreamFL, there is public data in the server on which the global model can be directly trained. However, we assume no training data on the server following the traditional FL setting. Therefore, we replace the centralized training on the server with an aggregation of the client models. We use 500 samples from the MS-COCO [2] dataset for knowledge distillation and set the optimal distillation and local contrastive weights as 1 and $1e-7$, respectively, after a parameter search.

**FedIoT [14].** We follow the original design of FedIoT by applying a factor of 100 to the multi-modal models during aggregation of the transformer blocks.

### B.2 Visualization of the data partitioning

We perform non-IID data partitioning to simulate the client data. For CIFAR-100 and AG NEWS datasets, we partition samples of each class with a random Dirichlet distribution with a given $\alpha$. For Flickr10k, we apply a non-IID number of training samples due to a lack of class labels, following [3]. A visualization of the number of samples on each client is shown in Fig. 2.

### B.3 Stochasticity discussion

To study the impact of the stochasticity in the experiments, we additionally conduct experiments with two additional random seeds besides the original seed 1 reported in the main paper and report the standard deviation (STD) of each method. As shown in Table 2, FedCola consistently outperforms all the comparison methods with a significant gap meanwhile holding the smallest STD, indicating the effectiveness and robustness of our proposed method.

### B.4 Breakdown performance

To provide more details for the reported results under each setting in the main paper, a breakdown performance with image-to-text top-1 recall (i2t $R$@1), text-to-image top-1 recall (t2i $R$@1) under both the 1k and 5k test image settings are given in Table 4 for Flickr and Table 5 for COCO Captions.

### B.5 Communication analysis

In §6.1 in the main paper, we study the communication trade-off of the proposed complementary local training. We further propose the communication costs of the comparison methods as a reference. Specifically, we report the size of the total download communication on one image client and one text client. An extended version is shown in Table 1. It shows that even FedCola with collaborative aggregation only (CA-only) can outperform all comparison methods without additional communication overhead. Further, when more communication budget is acceptable, FedCola (Attn) can provide a better trade-off between communication cost and performance, while the original FedCola can provide the highest performance.

**Table 1:** Communication cost and performance of each method on Flickr

| Method | Comm. Cost (MB) | $\mathbf{R@1_{sum}}$ |
|---|---|---|
| FedAvg | 208.81 | 81.08 |
| FedProx | 208.81 | 78.55 |
| CreamFL | 211.74 | 74.83 |
| FedIoT | 208.81 | 85.51 |
| FedCola (CA-only) | 208.81 | 90.09 |
| FedCola (Attn) | 262.95 | 91.73 |
| FedCola | 371.26 | 91.96 |

**Table 2:** Performance on Flickr under different random seeds. FedCola has the lowest standard deviation (STD).

| Method | Seed | | | STD ↓ |
|---|---|---|---|---|
| | 1 | 42 | 2024 | |
| FedAvg | 81.08 | 79.14 | 82.04 | 1.48 |
| FedProx | 78.55 | 77.86 | 81.69 | 2.04 |
| CreamFL | 74.83 | 75.94 | 78.34 | 1.79 |
| FedIoT | 85.51 | 80.16 | 81.10 | 2.86 |
| FedCola | **91.96** | **90.80** | **93.21** | **1.21** |

**Table 3:** Flickr performance under imbalanced total client numbers

| Setting | Method | 1k Test Image | | 5k Test Image | | $R@1_{sum}$ |
|---|---|---|---|---|---|---|
| | | i2t $R@1$ | t2i $R@1$ | i2t $R@1$ | t2i $R@1$ | |
| More Total Image Clients | FedAvg | 31.58 | 22.74 | 14.74 | 9.98 | 79.04 |
| | FedProx | 29.24 | 20.51 | 13.64 | 8.76 | 72.15 |
| | CreamFL | 29.58 | 21.34 | 13.84 | 9.22 | 73.98 |
| | FedIoT | 32.76 | 23.36 | 15.68 | 10.53 | 82.33 |
| | FedCola | **37.16** | **26.07** | **18.64** | **12.46** | **94.33** |
| More Total Text Clients | FedAvg | 32.90 | 23.34 | 15.48 | 10.39 | 82.11 |
| | FedProx | 20.02 | 14.60 | 7.88 | 5.64 | 48.14 |
| | CreamFL | 30.38 | 21.86 | 13.82 | 9.56 | 75.62 |
| | FedIoT | 31.88 | 22.85 | 14.82 | 10.29 | 79.84 |
| | FedCola | **36.24** | **25.76** | **17.62** | **12.06** | **91.68** |

### B.6 Imbalanced client scenario

In the main paper, we reported the performance when the number of participating clients is imbalanced and the number of total clients is the same as the default setting, considering the total client numbers in each type of client will only impact the uni-aggregation before the collaboration. To provide more experimental results, we report the performance under there are more image clients ($N_{\text{v}} = 16$ increased from 12) and more text clients ($N_{\text{l}} = 16$ increased from 12) in Table 3. As expected, FedCola still outperforms all comparison methods under such settings.

### B.7 Visualization

The smoothness of the parametric loss space has been utilized as a significant indicator of the model generalizabilty [1,8,12]. To illustrate that FedCola learns a more generalized global model, we visualize the loss space on 256 training samples of FedAvg (Fig. 3a) and FedCola (Fig. 3b) when the weights of the model are perturbed along the direction of the top Hessian eigenvectors. The loss landscape of FedCola is significantly smoother than FedAvg, indicating that with the help of the proposed framework, a more generalized global model can be obtained. Additionally, we further conduct visualizations with Linear Discriminant Analysis (LDA) at the feature level, as shown in Fig. 3c. By computing the distance between the feature centers, we find the gaps between uni-modal and multi-modal datasets are reduced under FedCola.

## C Potential Negative Societal Impact and Limitation

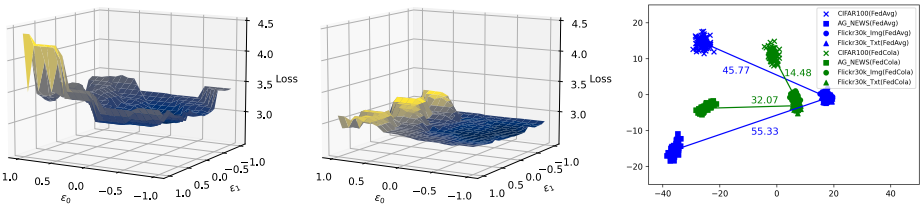
**Potential Negative Societal Impact.** The effectiveness of FedCola, like any machine learning model, is contingent on the data it's trained on. Given that data distribution in FL settings can be highly non-uniform and biased towards certain demographics or modalities, there's a risk of amplifying existing biases

**Table 4:** Flickr Breakdown Performance

| Setting | Method | 1k Test Image | | 5k Test Image | | $R@1_{sum}$ |
|---|---|---|---|---|---|---|
| | | i2t $R@1$ | t2i $R@1$ | i2t $R@1$ | t2i $R@1$ | |
| **Default** | FedAvg | 32.84 | 22.90 | 15.32 | 10.02 | 81.08 |
| | FedProx | 31.36 | 22.41 | 14.84 | 9.94 | 78.55 |
| | CreamFL | 30.2 | 21.34 | 13.82 | 9.46 | 74.83 |
| | FedIoT | 34.42 | 23.87 | 16.34 | 10.88 | 85.51 |
| | FedCola | **35.68** | **26.14** | **18.10** | **12.04** | **91.96** |
| **More Heterogeneity** | FedAvg | 32.5 | 23.34 | 15.40 | 10.46 | 81.70 |
| | FedProx | 31.1 | 22.06 | 13.9 | 9.26 | 76.33 |
| | CreamFL | 31.48 | 22.59 | 15.74 | 10.19 | 80.00 |
| | FedIoT | 33.02 | 23.73 | 15.94 | 10.57 | 83.28 |
| | FedCola | **36.26** | **26.06** | **17.54** | **11.96** | **91.82** |
| **Less Participation** | FedAvg | 25.84 | 18.95 | 11.96 | 8.06 | 64.82 |
| | FedProx | 25.94 | 19.02 | 10.74 | 7.64 | 63.33 |
| | CreamFL | 26.94 | 19.54 | 11.9 | 8.47 | 66.85 |
| | FedIoT | 25.18 | 18.13 | 11.02 | 7.62 | 61.94 |
| | FedCola | **34.94** | **25.48** | **16.60** | **11.84** | **88.85** |
| **More Image** | FedAvg | 31.22 | 22.68 | 14.42 | 9.96 | 78.28 |
| | FedProx | 31.46 | 22.84 | 14.90 | 10.05 | 79.25 |
| | CreamFL | 32.02 | 23.18 | 14.74 | 10.36 | 80.31 |
| | FedIoT | 33.22 | 23.40 | 15.78 | 10.34 | 82.74 |
| | FedCola | **35.42** | **25.8** | **17.76** | **12.26** | **91.24** |
| **More Text** | FedAvg | 31.94 | 22.55 | 15.20 | 10.00 | 79.69 |
| | FedProx | 31.20 | 22.25 | 14.44 | 9.70 | 77.59 |
| | CreamFL | 31.96 | 23.20 | 15.12 | 10.47 | 80.75 |
| | FedIoT | 31.46 | 22.22 | 14.56 | 9.77 | 78.02 |
| | FedCola | **35.48** | **25.50** | **17.40** | **11.72** | **90.10** |
| **Fewer Image-Text** | FedAvg | 24.92 | 18.01 | 10.70 | 7.49 | 61.12 |
| | FedProx | 24.28 | 17.50 | 10.22 | 7.19 | 59.19 |
| | CreamFL | 23.20 | 17.12 | 9.64 | 7.15 | 57.12 |
| | FedIoT | 24.68 | 17.76 | 10.42 | 7.47 | 60.34 |
| | FedCola | **34.06** | **24.28** | **16.18** | **11.16** | **85.68** |

**Table 5:** COCO Breakdown Performance

| Setting | Method | 1k Test Image | | 5k Test Image | | $R@1_{sum}$ |
|---|---|---|---|---|---|---|
| | | i2t $R@1$ | t2i $R@1$ | i2t $R@1$ | t2i $R@1$ | |
| **Default** | FedAvg | 36.98 | 29.28 | 16.76 | 12.40 | 95.42 |
| | FedProx | 37.56 | 28.46 | 16.68 | 12.46 | 95.16 |
| | CreamFL | 37.60 | 28.64 | 16.68 | 12.34 | 95.26 |
| | FedIoT | 38.62 | 29.97 | 17.16 | 12.65 | 98.40 |
| | FedCola | **41.02** | **31.62** | **18.74** | **13.72** | **105.10** |
| **More Heterogeneity** | FedAvg | 37.46 | 29.11 | 16.40 | 12.35 | 95.32 |
| | FedProx | 37.66 | 28.86 | 16.90 | 12.20 | 95.62 |
| | CreamFL | 35.76 | 28.11 | 15.74 | 11.80 | 91.41 |
| | FedIoT | 37.66 | 29.47 | 16.52 | 12.24 | 95.89 |
| | FedCola | **39.62** | **30.37** | **17.72** | **13.12** | **100.83** |
| **Less Participation** | FedAvg | 32.68 | 26.12 | 14.22 | 10.90 | 83.91 |
| | FedProx | 31.20 | 25.13 | 13.28 | 10.27 | 79.88 |
| | CreamFL | 31.58 | 25.00 | 12.60 | 9.79 | 78.97 |
| | FedIoT | 31.62 | 25.35 | 13.44 | 10.24 | 80.65 |
| | FedCola | **40.12** | **30.47** | **18.28** | **13.43** | **102.30** |
| **More Image** | FedAvg | 38.26 | 29.33 | 17.22 | 12.47 | 97.28 |
| | FedProx | 37.46 | 28.67 | 16.80 | 12.46 | 95.39 |
| | CreamFL | 36.80 | 28.66 | 16.02 | 12.17 | 93.65 |
| | FedIoT | 36.86 | 29.06 | 16.78 | 12.34 | 95.04 |
| | FedCola | **40.58** | **31.05** | **19.26** | **13.33** | **104.22** |
| **More Text** | FedAvg | 38.00 | 28.95 | 17.26 | 12.48 | 96.69 |
| | FedProx | 36.96 | 28.70 | 16.92 | 12.38 | 94.96 |
| | CreamFL | 36.68 | 28.46 | 15.62 | 12.06 | 92.81 |
| | FedIoT | 37.74 | 29.47 | 17.22 | 12.61 | 97.04 |
| | FedCola | **39.82** | **30.32** | **17.96** | **12.86** | **100.96** |
| **Fewer Image-Text** | FedAvg | 30.30 | 23.78 | 12.02 | 8.99 | 75.10 |
| | FedProx | 29.22 | 23.32 | 11.32 | 9.22 | 73.08 |
| | CreamFL | 29.60 | 23.78 | 12.18 | 9.12 | 74.69 |
| | FedIoT | 30.80 | 23.56 | 12.42 | 9.36 | 76.14 |
| | FedCola | **37.78** | **28.08** | **16.80** | **11.74** | **94.40** |




**(a)** FedAvg **(b)** FedCola **(c)** Extracted features

**Fig. 3:** Visualization of the parametric loss landscape with Hessian eigenvectors $\epsilon_0$ and $\epsilon_1$ and the extracted features for each resulting global multi-modal model.

or creating new ones. This can lead to unfair models that perform inequitably across different groups or modalities.

**Limitations.** FedCola currently does not address system heterogeneity, representing a limitation in the present framework. We propose to explore this aspect in future research.

## References

1. Chen, X., Hsieh, C.J.: Stabilizing differentiable architecture search via perturbation-based regularization. In: International conference on machine learn-

ing. pp. 1554–1565. PMLR (2020)
2. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
3. Hahn, S.J., Jeong, M., Lee, J.: Connecting low-loss subspace for personalized federated learning. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 505–515. KDD '22, Association for Computing Machinery, New York, NY, USA (2022)
4. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International conference on machine learning. pp. 5132–5143. PMLR (2020)
5. Krizhevsky, A.: Learning multiple layers of features from tiny images pp. 32–33 (2009)
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
7. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282 (2017)
8. Mendieta, M., Yang, T., Wang, P., Lee, M., Ding, Z., Chen, C.: Local learning matters: Rethinking data heterogeneity in federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8397–8406 (2022)
9. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
10. Potamianos, G., Marcheret, E., Mroueh, Y., Goel, V., Koumbaroulis, A., Vartholomaios, A., Thermos, S.: Audio and visual modality combination in speech processing applications. In: The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations-Volume 1, pp. 489–543 (2017)
11. Yu, Q., Liu, Y., Wang, Y., Xu, K., Liu, J.: Multimodal federated learning via contrastive representation ensemble. In: The Eleventh International Conference on Learning Representations (2022)
12. Zela, A., Elsken, T., Saikia, T., Marrakchi, Y., Brox, T., Hutter, F.: Understanding and robustifying differentiable architecture search. In: International Conference on Learning Representations (2020)
13. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. Advances in neural information processing systems **28** (2015)
14. Zhao, Y., Barnaghi, P., Haddadi, H.: Multimodal federated learning on iot data. In: 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI). pp. 43–54 (2022)