Breaking Barriers of System Heterogeneity: Straggler-Tolerant Multimodal Federated Learning via Knowledge Distillation

Jinqian Chen[1,3], **Haoyu Tang**[1*], **Junhao Cheng**[1], **Ming Yan**[2], **Ji Zhang**[2], **Mingzhu Xu**[1], **Yupeng Hu**[1] and **Liqiang Nie**[4]
[1]School of Software, Shandong University
[2]Alibaba Group
[3]School of Software Engineering, Xi'an Jiaotong University
[4]Harbin Institute of Technology (Shenzhen)
j1nqianchen6@gmail.com, {tanghao258, lord.c, xumingzhu, huyupeng}@sdu.edu.cn, {ym119608, zj122146}@alibaba-inc.com, nieliqiang@gmail.com

Abstract

Internet of Things (IoT) devices possess valuable yet private multimodal data, calling for a decentralized machine learning scheme. Though several multimodal federated learning (MFL) methods have been proposed, most of them merely overlook the system heterogeneity across IoT devices, resulting in the inadaptability to real world applications. Aiming at this, we conduct theoretical analysis and exploration experiments on straggler impacts and uncover the fact that stragglers caused by system heterogeneity are fatal to MFL, resulting in catastrophic time overhead. Motivated by this, we propose a novel Multimodal Federated Learning with Accelerated Knowledge Distillation (MFL-AKD) framework, which is the first attempt to integrate knowledge distillation to combat stragglers in complex multimodal federated scenarios. Concretely, given the pretrained large-scale vision-language models deployed in the central server, we apply a fast knowledge transfer mechanism to conduct early training of local models with part of the local data. The early-trained model is then enhanced through the distillation of the pretrained large model and further trained on the remaining data. Extensive experiments on two datasets for video moment retrieval and two datasets for image-text retrieval demonstrate that our method achieves superior results with high straggler robustness.

1 Introduction

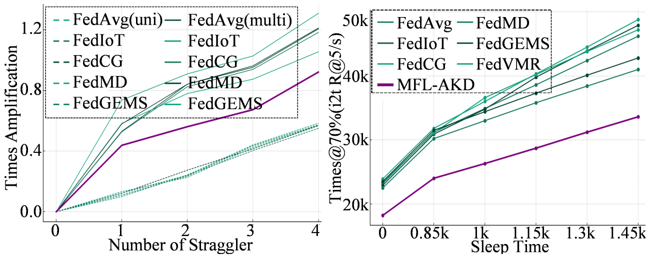
Nowadays, with the increase of multimedia data (e.g., images, videos, and texts) in daily life, the imperative to harness the wealth of multimedia data has become a hot topic, which has raised increasing research interests in many multimodal tasks such as text-image retrieval [Qu *et al.*, 2021] and video moment retrieval [Gao *et al.*, 2017; Wang *et al.*, 2021]. The integration and storage of these data modalities across mobile and IoT systems present a significant challenge in training models with data from diverse devices.

Multimodal Federated Learning (MFL) [McMahan *et al.*, 2017; Yang *et al.*, 2020] has emerged as a promising approach to leverage multimodal data from various sources such as enormous IoT devices without privacy disclosure. These IoT devices always vary in computation capacity, communication bandwidth, energy power, and operation systems [Han *et al.*, 2020]. All these factors lead to so-called system heterogeneity. However, existing MFL methods have overlooked this intrinsic property with a naive assumption: uniform capacities and the same model structures across all clients [Cheng *et al.*, 2021; Wang *et al.*, 2021]. Such an ideal assumption eliminates the heterogeneous nature of MFL, neglecting the severe time overhead and performance degradation caused by the stragglers.

Stragglers, or less efficient participants, have been recognized as a fundamental challenge in FL since its inception [Wu *et al.*, 2020]. The huge extra time overhead with the performance degradation [Kairouz *et al.*, 2021] caused by stragglers have led to the development of various mitigation methods, which can be broadly categorized into two groups: (1) employing relaxed synchronization, (2) improved scheduling and aggregation schemes. What's worse, almost all these methods cannot be applied to MFL directly. For synchronization-based methods, the outdated models and dropped knowledge parts are deadly to multimodal scenarios. Additionally, aggregation-based methods face difficulties due to the expected diversity in model structures among clients, a result of the inherent system variability in MFL. This variation makes it challenging to directly combine local gradients, rendering the aggregation approach impractical in MFL.

Focusing on the issue, we carefully investigate the practical impact of stragglers in MFL under the real-world scenarios. Instead of simulation, we conduct all the explored experiments using real decentralized machines with different controlled computation capacities and communication bandwidth. As shown in Fig. 1, our findings reveal that the impact of stragglers in MFL is more severe than in horizontal single-modal FL, leading to significant delays and even training failure, and this problem is exacerbated as the number

*Corresponding Author: Haoyu Tang



(a) Time increase v.s stragglers. (b) Times@70% v.s sleep time

Figure 1: Illustrations of Straggler issues. In left figure, a straggler is set to 50% reduced speed compared to a standard client. 'Uni' and 'multi' represent unimodal and multimodal FL, respectively. The right figure shows the results from MSCOCO for i2t retrieval.

and latency of stragglers increase. All these findings suggest a significant implication: incomprehensive aggregation of cross-modal information is fatal to multimodal decentralized learning, which calls for the development of a straggler-robust MFL framework under practical system heterogeneity.

Motivated by these observations, we decompose the challenge and reorganize them into two requirements: (1) Efficiency and lightness for locally deployed models, ensuring minimal computation and communication demands, paired with a powerful, large-scale global model for in-depth cross-modal understanding. (2) An effective, adaptable aggregation strategy, not only enabling knowledge obtainment from different model structures, but also allowing partial aggregation to avoid total falls while maintaining the possibility of full aggregation. To this end, we propose a multimodal federated learning framework with accelerated knowledge distillation (MFL-AKD). To alleviate the local budget while keeping the powerful representative ability of the central server, we deploy a lightweight encoder to each client while maintaining the powerful CLIP in the central server. For flexible aggregation and potential stragglers, we incorporate principles of knowledge distillation for efficient knowledge transfer from the central model to local models, and designed a fast knowledge transfer mechanism before the regular updating in each round, accelerating the convergence of the federated models. We validate the proposed MFL-AKD on two crucial vision-language tasks, text-image retrieval [Qu *et al.*, 2021] and video moment retrieval [Tang *et al.*, 2021a; Tang *et al.*, 2021b; Hu *et al.*, 2023], on four popular datasets. Experimental results validate the superior performance and the remarkable straggler-tolerant ability of our proposed MFL-AKD. The main contributions of this paper are summarized as follows:

- We take the first attempt to identify and address the straggler problem in multimodal federated learning. Through analysis and exploration experiments with real decentralized machines, we discovered that stragglers affect MFL more severely than single-modal FL, significantly impacting the central model's effectiveness.
- We decompose the straggler challenge in MFL and propose Multimodal Federated Learning with Accelerated Knowledge Distillation. To mitigate potential straggler issues, MFL-AKD introduces a Fast Knowledge Transfer Mechanism to circumvent potential setbacks caused by large-scale data and provide guidance for subsequent full-batch training, effectively combating the stragglers with its fast convergence and early-training strategy.
- We demonstrate the effectiveness of our approach on two crucial vision-language tasks: text-image retrieval and video moment retrieval. Extensive experiments confirm that our method outperforms existing MFL approaches, demonstrating superior performance and remarkable tolerance to straggler-related delays.

2 Related Work

2.1 Stragglers in Federated Learning

In recent years, federated learning has gained increasing attention for its ability to train deep models using decentralized data on the client side, without sharing raw local data [Li *et al.*, 2020a; Kairouz *et al.*, 2021]. However, system heterogeneity leads to stragglers that can significantly slow down the convergence rate of synchronous FL algorithms, posing a serious issue for real-time applications. To address this, researchers have proposed asynchronous FL algorithms to mitigate straggler lag by relaxing client synchronization requirements [Wu *et al.*, 2020; Liu *et al.*, 2021]. However, these methods face challenges with communication bottlenecks and outdated local model updates during aggregation. Therefore, current research efforts primarily focus on straggler issues in synchronous FL. Common strategies involve setting fixed deadlines for all clients to update and share their local models, allowing clients to process data samples at varying speeds based on computational capacity [Smith *et al.*, 2017; Li *et al.*, 2020b; Han *et al.*, 2020]. Additionally, [Reisizadeh *et al.*, 2022] proposed to first train with faster clients and then gradually include stragglers towards the end of training. However, these approaches have not investigated the MFL scenario, which is more severely impacted by stragglers. They also lack a real-time update strategy in federated learning knowledge distillation environments, which is crucial for maintaining efficiency and effectiveness in MFL with multiple stragglers and extended delays.

2.2 Knowledge Distillation

Knowledge distillation (KD) aims to transfer the knowledge of a large and complex model (i.e., the teacher model) to a smaller and simpler model (i.e., the student model) so that a lightweight student with comparable performance to the teacher can be obtained [Hinton *et al.*, 2015; Romero *et al.*, 2015]. This technique often minimizes the discrepancy in probability distributions of teacher and student models between their final logical outputs [Hinton *et al.*, 2015] or intermediate features [Komodakis and Pesquet, 2017].

Recently, knowledge distillation has demonstrated encouraging outcomes in compressing the client model size of unimodal FL [Huang *et al.*, 2022]. For example, [Ahn *et al.*, 2020] presented to transfer the knowledge from a global model to local models with a weighted loss function to balance the knowledge distillation and the federated learning objectives. [Li *et al.*, 2020c] proposed a FedMD framework that integrates transfer learning and knowledge distillation to facilitate federated learning in scenarios where individual clients possess their distinctive model designs.

Despite the progress they have made, directly extending those unimodal knowledge distillation methods into the multimodal FL scenarios will be highly inappropriate due to the heterogeneity of cross-modal data. More importantly, the straggler problem not addressed by these federated knowledge distillation methods will lead to more severe performance degradation and cost increases [Bonawitz *et al.*, 2017].

2.3 Multimodal Federated Learning

Multimodal federated learning is an emerging area that involves using various data sources (e.g., text, image, and audio) from multiple clients to develop a multimodal model, ensuring data privacy across different clients. To address this issue, several methods have been proposed in recent years [Cheng *et al.*, 2021; Wu *et al.*, 2022; Li *et al.*, 2020c; Zhao *et al.*, 2021]. For example, [Liu *et al.*, 2019] introduced FL to the vision-and-language grounding tasks and proposed to collaboratively extract diverse image representations derived from different tasks. Inspired by the idea of contrastive learning, [Huang *et al.*, 2022] presented a CreamFL framework that regularizes local client training by incorporating both inter-modal and intra-modal contrasts to enhance the multimodal FL. Besides, [Wang *et al.*, 2021] aimed at the video moment retrieval and proposed to attentively aggregate the model of grouped clients that are trained sequentially.

However, as these methods do not involve knowledge distillation and specialized acceleration strategy design, all of them will face severe performance degradation and communication delays when facing stragglers. In contrast, our MFL-AKD ensures its robustness to stragglers through the designed quick knowledge transfer strategy in KD process.

3 Revisiting Stragglers in Synchronized FL

To investigate the difference in the impact of stragglers in unimodal FL and multimodal FL, we here theoretically revisit and further conduct exploration experiments.

3.1 Problem Setup

Consider a multimodal federated learning scenarios with K clients with up to N modals data. Each client c_i possesses n_i samples, constituting the local dataset $\mathcal{D}_i = \{x_i^j, y_i^j\}_{j=1}^{n_i}$. The goal of multimodal federated learning is to collaboratively train a global model $f(\cdot;\theta_g)$ parameterized by θ_g by utilizing the decentralized datasets $\{\mathcal{D}_i\}_i^K$ in T communication rounds. At the beginning of each round, the server will decide the participated clients set $\mathbb{P}$. Let $\mathbb{I}^t = \{i|c_i \in P\}$. The participant number in the t-th round is denoted as $\tau^t = |\mathbb{P}^t| = |\mathbb{I}^t|$. Each client $c_i \in \mathbb{P}$ is then required to conduct the training task on its local data $\mathcal{D}_i$. Considering the system heterogeneity across $\mathbb{P}^t$, there may exist M^t stragglers in the t-th communication round, and $M^t \leq \tau^t$. Unimodal FL is the special case when $N = 1$ under this definition.

3.2 The Impact of Stragglers

System heterogeneity is an inevitable issue that occurs in the practical application of FL. Due to the difference in hardware, e.g. computing chips, battery power, and communication bandwidth, the ability to execute training tasks varies across different clients, which causes the well-known stragglers problem. Though numerous methods have been elaborated to deal with such an issue, the theoretical analysis of straggler impact in FL is yet to be established.

Definition 1 *(Time Consumption of Synchronized FL) Consider a synchronized federated framework with K clients $\{c_i\}_{i=1}^K$ for R communication rounds, the server requires each participated client $c_i \in \mathbb{P}^r$ to conduct the training task on its local dataset $\mathcal{D}_i$. Let ϵ_i^r denote the total floating point operation number of client c_i in the r-th round, γ_i^r denote the comprehensive computation capacity measured by floating point operations per second (FLOPS). The time consumption of FL is defined as:*

$$\Upsilon = \sum_{r=1}^{R} \max\left(\left\{\left(\frac{\epsilon_i^r}{\gamma_i^r} + \zeta_i^r\right)|i \in \mathbb{I}^r\right\}\right) \tag{1}$$

where $\max()$ is a function to extract the max element in a set and ζ_i^r is the communication time cost of client c_i in the r-th round.

To further facilitate the numerical analysis, we introduce the sum of the averaged wasted time across all clients during a full federated training procedure as a metric to measure the absolute impact of stragglers in FL defined as below:

Definition 2 *(Averaged Wasted Time of Stragglers) Given a federated framework with its relevant variables as introduced before, the averaged wasted time of stragglers is defined as:*

$$\delta = \sum_{r=1}^{R}\sum_{i\in\mathbb{I}^r} \frac{1}{|\mathbb{I}^r|}\left|\frac{\epsilon_i^r}{\gamma_i^r} + \zeta_i^r - \upsilon\right| \tag{2}$$

where $\upsilon = \max\left(\left\{\left(\frac{\epsilon_i^r}{\gamma_i^r} + \zeta_i^r\right)|i \in \mathbb{I}^r\right\}\right)$ is the time consumption of the lowest client in r-th round.

We now claim our analysis of straggler impacts as indicated in Proposition 1.

Proposition 1 *The impact of stragglers is correlated with: (1) Task Difficulty Φ; (2) Balance of dataload on unit computation capacity; (3) Communication consumption.*

Analysis (1) Task difficulty Φ closely relates to the required communication rounds R for desired model performance. The convergence speed within a federated algorithm is influenced by the convexity and smoothness of the objective functions. Therefore, simpler tasks that satisfy the conditions of convexity and smoothness, tend to converge more rapidly. (2) Workload on unit computation capacity is decided by the data amount and the model capacity. The more imbalance the dataload on unit computation, the more gap exists in the time consumption across all clients.

Discussion (1) The common trick to mitigate severe stragglers involves setting a response threshold, effectively reducing υ to lessen the time consumption gap $\left|\frac{\epsilon_i^r}{\gamma_i^r} + \zeta_i^r - \upsilon\right|$. (2) The communication budget is generally constant, as the uploaded content consistently includes the gradient or parameters of the current local model. Therefore, communication

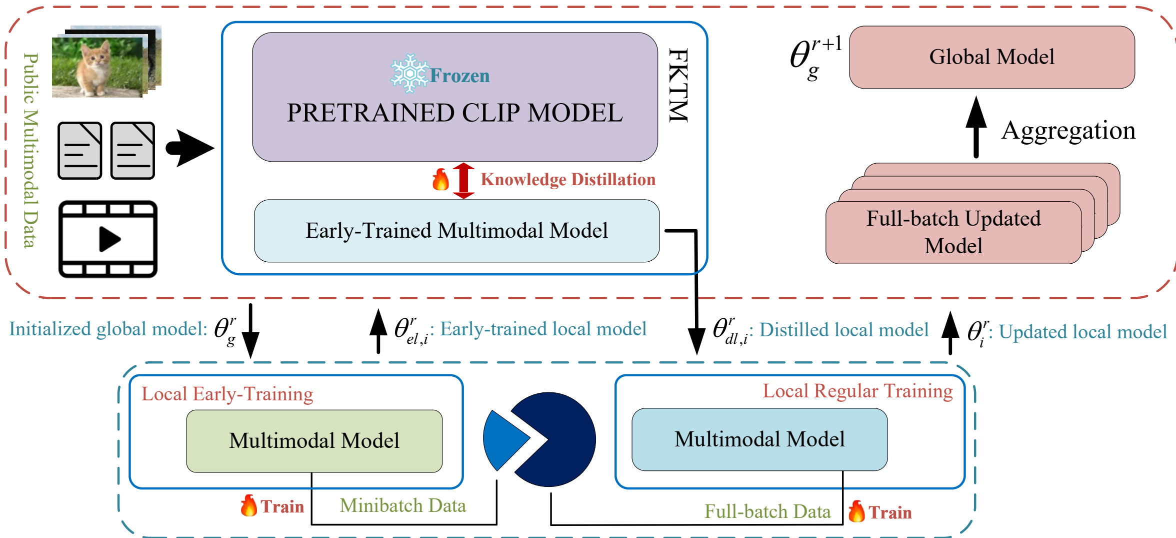


Figure 2: An illustration of the proposed MFL-AKD framework, which consists of the following steps in each communication round: 1) distribute the initialized global model θ_g^r to all participated clients; 2) early training is conducted on the part of local data; 3) upload early-trained model $\theta_{\mathrm{el},i}^r$ to the server; 4) distill knowledge from pretrained vision-language model to get distilled local model $\theta_{\mathrm{dl},i}^r$; 5) re-distribute $\theta_{\mathrm{dl},i}^r$ to its client and conduct regular training; 6) aggregation to get global model θ_g^{r+1}.

consumption ζ_i mainly depends on the communication bandwidth, which is hard to manually optimize in federated algorithm design.

Comparison of multimodal FL with unimodal FL: (1) Tasks in multimodal FL are often more difficult than those in unimodal FL, as they involve the joint relationships between different modalities. (2) Workload is more imbalanced across clients in multimodal FL than in unimodal FL, as data in different modalities possesses different data structures with different storage requirements. Those differences magnify the impact of stragglers in multimodal FL, calling for the establishment of a straggler-robust multimodal FL framework.

4 The Proposed Method

In this section, we introduce our proposed straggler-robust multimodal federated framework MFL-AKD. We choose the vision-language tasks as representative multimodal tasks. Specifically, we consider text-to-image and video moment retrieval tasks, which aim to match a given sentence x with its most relevant image or video moment y.

4.1 Motivation

We now briefly revisit our motivation for the proposal of MFL-AKD. As described before, system heterogeneity is an inevitable issue in the practical application of multimodal federated learning, causing stragglers in each communication round. However, such an issue has been largely overlooked with a naive conjecture that the impact of stragglers in MFL is similar to that in the unimodal FL. With the analysis proposed in Section 3.2, the impact of stragglers in multimodal FL is much more severe than that in unimodal FL, which calls for the establishment of a straggler-robust multimodal federated framework. To develop such a framework, the following requirements should be satisfied: From the system perspective: (1) Lightweight but effective client-deployed model to alleviate the workload of clients. (2) Powerful and large server-deployed model to sufficiently handle the difficult multimodal tasks. From the scheduling perspective: (1) Balanced workload decided by computation capacities across all clients; (2) Efficient aggregation schema to accelerate the convergence.

4.2 Framework Overview

In response to these requirements, we develop a straggler robust multimodal federated method named MFL-AKD. To handle the challenge brought by multimodal tasks, MFL-AKD deploys a powerful pretrained vision-language model CLIP $\mathrm{F}(\cdot;\theta_C)$ on the server and allows clients to use its desired model $f_i(\cdot;\theta_i)$ in arbitrary structures. To facilitate the convergence of the federated algorithm and allow the dynamic workload to prevent straggling, we have designed a novel Fast Knowledge Transfer Mechanism (FKTM) in MFL-AKD. FKTM requires all clients to extract part of its local data to conduct workload-balanced early training before the ordinary training. Such a partial extraction ensures the balance of workload on computation capacity, largely reducing the wasted time of stragglers. The early-trained models are then uploaded to the server to conduct a warm-up utilizing the CLIP model through knowledge distillation. By doing this, FKTM not only mitigates the time consumption gap between clients but also accelerates the convergence of the federated framework. The application of KD also allows the deployment of models in different structures on clients. Note that we here mainly introduce MFL-AKD with the same model architectures deployed in clients to facilitate the comparison with other MFL methods. The overall framework of MFL-

AKD is illustrated in Figure 2.

Commencing the r-th round, the central server first identifies the participated client set $\mathbb{P}^r$. For each client $c_i \in \mathbb{P}^r$, the server distributes current global model parameter θ_g^r to initialize its local model. Concerning the possible imbalance workload on computation capacity $\frac{\epsilon_i^r}{\gamma_i^r}$ among clients set $\mathbb{P}^r$, MFL-AKD requires each client $c_i \in \mathbb{P}^r$ to randomly extract a mini-batch of local data D_i according to its computation capacity ϵ_i^r in current round and further conduct early-training to get the local model parameter $\theta_{\text{el},i}^r$, which is further asynchronously uploaded to the server. Receiving the uploaded early-trained parameter $\theta_{\text{el},i}^r$ of client c_i, the FKTM in the server then utilizes the CLIP model to conduct knowledge distillation to warm up the early-trained local model, accelerating the convergence of the federated frameworks. The distilled early-trained local model $\theta_{\text{dl},i}^r$ is then distributed to its corresponding client c_i and further conduct model training on the rest data to get updated local model θ_i^r. All the updated local model $\{\theta_i^r\}_{i\in\mathbb{I}^r}$ is then uploaded to the server to get global model θ_g^{r+1} for next round training.

4.3 Multimodal Local Training

Without loss of generality, we introduce the local training of MFL-AKD under the multimodal retrieval tasks. The input instance $\{x_i, y_i\}$ in the local dataset $\mathcal{D}_i$ denotes a sentence x_i and its related image or video moment y_i, and the training goal is to obtain a local multimodal encoder $f_i(\cdot;\theta_i^r)$ that characterizes the visual language well. Considering the excellence of transformer in cross-modal embedding, we adopted a transformer-structured encoder [Vaswani *et al.*, 2017].

Specifically, f_i encodes x_i and y_i into their respective embeddings h_{x_i} and h_{y_i}. The local loss function that minimizes the distance between h_{x_i} and h_{y_i} is defined as:

$$\min_{f_i} \frac{1}{n_i} \sum_{n=1}^{n_i} L(h_{x_i}, h_{y_i}) \tag{3}$$

where n_i denotes the size of $\mathcal{D}_i$, and L denotes the mean square error.

4.4 Fast Knowledge Transfer Mechanism (FKTM) via Centralized Knowledge Distillation

We here specifically introduce the designed FKTM with a centralized knowledge distillation process in MFL-AKD. As analyzed before, the impact of stragglers on a particular federated framework is highly correlated with its convergence speed and balance of workload on unit computation capacity across clients. To effectively handle the straggler problem and prevent entirely dropped clients, FKTM requires the clients $c_i \in \mathbb{P}^r$ to early-train the initialized local model θ_g^r on a small fraction of its local data while remaining the workload balance. After receiving the local early-trained model $\theta_{\text{el},i}^r$, FKTM further enhances the ability of the model by distilling knowledge from the powerful teacher model. Specifically, a large pretrained CLIP model $\mathrm{F}(\cdot;\theta_C)$ is adopted as the teacher. For the input pair (x_i^j, y_i^j), the CLIP teacher and the student encoder $\theta_{\text{el},i}^r$ embed them into their corresponding representations and further conduct knowledge distillation through:

$$\mathcal{L}_{\text{dist}} = \mathrm{KL}\left(F(x_i^j \mid \theta_C) \| f(x_i^j \mid \theta_{\text{el},i}^r)\right) \tag{4}$$

The distilled local model $\theta_{\text{dl},i}^r$ of client c_i is then redistributed to the client to continue the training on the rest of its local data. FKTM sufficiently utilizes the prior knowledge in pretrained large multimodal models to help the early-trained model rapidly adapt to its local multimodal data and get experience from the prior knowledge. Such a mechanism significantly accelerates the convergence of the MFL-AKD, making the framework tolerant to stragglers and leaving readiness models for each client to combat the potential drop-offs during the subsequent training.

5 Experiments

We evaluated our framework on two fundamental vision-language tasks: image-text retrieval on the Flickr30k and MS COCO datasets, and video moment retrieval on the Charades-STA and ActivityNet Captions datasets.

5.1 Experimental Setup

Datasets We use four popular multimodal datasets in text-image and text-video retrieval tasks. Details are given below:

1. **Flickr30k** [Young *et al.*, 2014]: This image-text retrieval dataset consists of 31,784 images, each of which is manually annotated with five different sentence descriptions. As in [Qu *et al.*, 2021], 29,784, 1,000, and 1,000 images with paired sentences are adopted for training, validation, and testing, respectively.
2. **MSCOCO** [Lin *et al.*, 2014]: This image-text retrieval dataset contains 123,287 images, each of which is paired with five annotated sentences. For fair comparisons, the public dataset split is adopted [Qu *et al.*, 2021], i.e., 113,287, 5000, and 5000 images for training, validation, and testing, respectively. Besides, the **MSCOCO 5Fold 1K** setting is adopted for evaluation, where the average results are over 5-fold of 1,000 testing images.
3. **Charades-STA** [Gao *et al.*, 2017]: This dataset video moment retrieval is manually annotated by [Gao *et al.*, 2017], which contains 6,672 videos with 29.76 seconds long on average. The number of sentence-video pairs is 16,127 in total. Following the common settings [Gao *et al.*, 2017], we divide those pairs into two parts, i.e., 12408 pairs for training and 3720 pairs for testing, respectively.
4. **ActivityNet Captions** (Anet) [Krishna *et al.*, 2017]: This video moment retrieval dataset contains 14926 videos with an average duration of 120 seconds. The sentence-video pairs are 71,957 in total, where the corresponding sentences are longer with more complicated semantics. Following [Gao *et al.*, 2017], we adopt 37,417, 17,505, and 17,031 sentence-video pairs for training, validation, and testing, respectively.

Evaluation metrics. The standard Recall@K (R@K for short) and R@n, IoU=m are adopted as the evaluation metrics for the image-text retrieval and video moment retrieval

Method	Flickr30k				MSCOCO 5Fold 1K			
	Text Retrieval		Image Retrieval		Text Retrieval		Image Retrieval	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
FedAvg[McMahan *et al.*, 2017]	67.9	89.0	47.8	73.1	64.8	89.8	46.6	79.6
FedIoT[Zhao *et al.*, 2021]	67.2	88.7	45.4	71.9	64.0	89.6	43.7	79.1
FedCG[Wu *et al.*, 2022]	66.5	87.8	44.9	70.4	63.8	88.5	45.8	78.7
FedMD[Li *et al.*, 2020c]	67.4	89.1	47.9	73.8	64.9	90.7	46.4	80.3
FedGEMS[Cheng *et al.*, 2021]	**69.4**	90.7	50.9	74.5	66.0	<u>94.2</u>	54.0	82.3
FedVMR[Wang *et al.*, 2021]	<u>69.3</u>	<u>91.3</u>	<u>51.2</u>	<u>75.2</u>	**66.3**	94.1	<u>54.2</u>	<u>82.8</u>
MFL-AKD (w/o. KD)	66.8	88.4	46.9	72.3	64.8	92.9	53.3	81.5
MFL-AKD (w/o. FKTM)	68.3	89.6	49.2	72.1	66.0	92.7	53.4	81.7
MFL-AKD	69.0	**91.5**	**51.3**	**75.8**	<u>66.1</u>	**94.2**	**54.4**	**83.1**

Table 1: Performance comparison of Text-Image Retrieval on Flickr30k and MSCOCO datasets (%).

Models	Charades-STA				Anet			
	R@1		R@5		R@1		R@5	
	n=0.5	n=0.7	n=0.5	n=0.7	n=0.5	n=0.7	n=0.5	n=0.7
FedAvg[McMahan *et al.*, 2017]	50.5	25.5	82.4	58.3	36.9	17.6	78.9	63.7
FedIoT[Zhao *et al.*, 2021]	47.9	23.6	80.4	57.2	37.8	18.6	79.3	64.1
FedCG[Wu *et al.*, 2022]	39.4	19.7	78.9	53.5	20.0	4.2	71.4	58.3
FedMD[Li *et al.*, 2020c]	51.7	26.4	82.9	59.1	39.4	19.8	80.4	64.3
FedGEMS[Cheng *et al.*, 2021]	52.9	<u>31.7</u>	83.6	64.4	41.7	22.5	81.2	65.3
FedVMR[Wang *et al.*, 2021]	<u>53.0</u>	31.6	**84.7**	<u>64.7</u>	<u>42.3</u>	<u>22.7</u>	**82.6**	**66.2**
MFL-AKD (w/o. KD)	49.3	25.4	79.2	53.6	34.8	16.5	79.0	61.2
MFL-AKD (w/o. FKTM)	49.2	24.9	80.1	53.4	36.9	20.1	80.6	64.1
MFL-AKD	**53.1**	**32.2**	<u>84.6</u>	**65.0**	**42.4**	**23.4**	<u>82.4</u>	<u>66.1</u>

Table 2: Performance comparison of Video Moment Retrieval on Charades-STA and Anet datasets (%).

task, respectively. R@K denotes the percentage of ground truth at the top-K retrieved results, and R@n, IoU=m means the percentage of the samples that contain at least one out of top-n retrieval results with Intersection over Union (IoU) larger than m. Besides, we also evaluate the efficiency of our framework by measuring its relative time gain [Wu *et al.*, 2022] compared to a predefined baseline.

Compared methods. We compare the MFL-AKD with various state-of-the-art (SOTA) multimodal FL methods to validate its superior performance and straggler robustness, including FedAvg [McMahan *et al.*, 2017], FedIoT [Zhao *et al.*, 2021], FedCG [Wu *et al.*, 2022], FedMD [Li *et al.*, 2020c], FedGEMS [Cheng *et al.*, 2021], and FedVMR [Wang *et al.*, 2021]. Among those methods, FedIoT, FedCG, and FedGEMS are multimodal FL frameworks, where we conduct experiments on four datasets and report the corresponding results. and FedMD is an unimodal method with knowledge distillation, where we extend this method to accommodate the multimodal settings.

Implementation details. Based on our MFL-AKD framework, we conducted experiments with 40 to 60 communication rounds for the federated learning process. All experiments were performed on a cluster of 4 heterogeneous devices with different configurations. We implemented our framework using PyTorch 1.7.1. For the server side, we used a high-performance computing node equipped with four Intel Xeon processors and 128GB memory. For the vision-language knowledge distillation, the ViT-B/32 version CLIP [Radford *et al.*, 2021] is adopted as the teacher model. Each client device was equipped with a GPU of a different model and memory size. We randomly partitioned the datasets among the clients. For the text-image retrieval task, the model on each client is trained locally for 10 rounds and 30 rounds on the Flicker30k and MSCOCO datasets, respectively. For the video moment retrieval task, the model on each client is trained for 40 rounds and 60 rounds on the Charades-STA and Anet datasets, respectively. The stochastic gradient descent (SGD) optimizer with a learning rate of 0.001 is adopted. During the federated learning process, we applied knowledge distillation with a temperature of 5 and a weight of 0.5 to encourage model convergence.

5.2 Performance Comparison

We first compare the performance of MFL-AKD with SOTA multimodal FL methods on four multimodal datasets. Table 1 and Table 2 display the retrieval performance of MFL-AKD and other baselines for video moment retrieval and text-image retrieval tasks, respectively. Note that the best results are highlighted and the second ones are underlined.

From those results, the following observations stand out.

Times@70% (i2t R@5/s)	$\upsilon = 1500$	1750	2000	2250	2500
FedAvg	20.2k	22.8k	25.3k	28.8k	31.3k
FedIoT	20.1k	22.0k	24.2k	26.3k	28.6k
FedCG	20.0k	22.4k	24.9k	27.3k	29.8k
FedMD	19.9k	22.3k	24.8k	27.1k	29.5k
FedGEMS	19.1k	20.8k	22.6k	24.5k	26.3k
FedVMR	19.8k	22.0k	24.3k	26.6k	28.9k
MFL-AKD (w/o. KD)	19.6k	21.7k	23.8k	26.0k	28.2k
MFL-AKD (w/o. FKTM)	17.4k	19.2k	21.0k	22.7k	24.2k
MFL-AKD	**15.6k**	**16.8k**	**17.9k**	**19.1k**	**20.3k**

Table 3: Robustness to Stragglers of Text-Image Retrieval Tasks on Flickr30k Dataset with different straggling time υ.

Times@50% (R@5 IoU=0.7/s)	$\upsilon = 750$	900	1050	1200	1350
FedAvg	37.4k	42.2k	47.0k	51.7k	56.3k
FedIoT	37.0k	41.7k	46.2k	50.6k	54.8k
FedCG	38.1k	42.8k	48.5k	52.8k	57.9k
FedMD	36.9k	41.0k	45.2k	49.2k	53.3k
FedGEMS	36.1k	40.0k	43.8k	47.5k	51.2k
FedVMR	35.5k	39.1k	42.7k	46.3k	49.8k
MFL-AKD (w/o. KD)	33.6k	37.0k	40.5k	43.9k	47.4k
MFL-AKD (w/o. FKTM)	32.9k	36.5k	50.0k	43.2k	46.8k
MFL-AKD	**30.1k**	**33.0k**	**35.5k**	**37.9k**	**40.2k**

Table 4: Robustness to Stragglers of Video Moment Retrieval Tasks on Anet Dataset with different straggling time υ.

For the text-image retrieval task, the proposed MFL-AKD model consistently surpasses all baselines over most metrics. As for "R@1" metric on both datasets, we also achieve a competing performance (69.0 vs 69.4). Furthermore, our MFL-AKD model also achieves the best retrieval performance in terms of all metrics except for "R@5, IoU=0.5" on both datasets. Compared to the strongest FedVMR baseline which targets this task, our method achieves substantial improvements on most metrics.

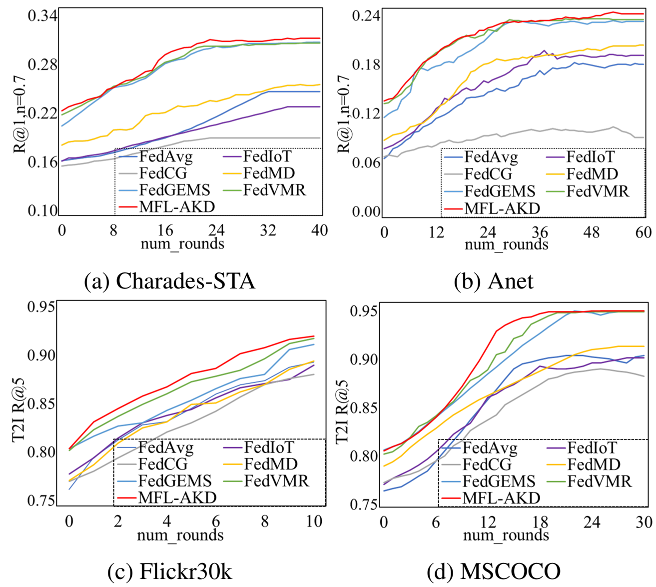
We further compare the convergence speed of all federated methods on various multimodal benchmarks. Results in Fig. 3 demonstrate that MFL-AKD converges fast on all benchmarks, significantly surpassing all compared methods.

5.3 Robustness on Stragglers

We further conduct experiments to validate the effectiveness of MFL-AKD in combating the stragglers in MFL scenarios. To measure the severity of the straggler problems, we adopt the same implementation strategy with the performance comparison experiments, while manually changing the straggling time. We use the time consumption Υ to compare the robustness of stragglers among various methods. Table 3 and Table 4 report the experimental results on Flickr30k for image-text retrieval tasks and Anet for video moment retrieval tasks respectively. As demonstrated in the table, MFL-AKD exhibits the shortest time consumption under different straggling times, and when the straggling time varies from the lowest to the highest, the time consumption only increases by around 5k for text-image retrieval and 10k for video moment retrieval, which is far lower than that of other baselines.

5.4 Ablation Studies

To evaluate the contribution of different components in our proposed MFL-AKD framework, we further conduct ablation experiments to validate the effectiveness of the design



(a) Charades-STA (b) Anet

(c) Flickr30k (d) MSCOCO

Figure 3: Fast Convergence of MFL-AKD on benchmarks.

of MFL-AKD. Specifically, we gradually remove key components of our framework and obtained the following model variants.

- MFL-AKD (w/o. FKTM): We discard the FKTM from MFL-AKD which disables the early training and adopt the common communication strategy as FedAvg. The knowledge distillation of pretrained model is now conducted on the global updated model after the aggregation of the clients' model.
- MFL-AKD (w/o. KD): We remove the knowledge distillation mechanism from MFL-AKD, which directly returns the aggregated model of early-trained models without learning from pretrained large model.

The results are illustrated in the corresponding Table 1, 2, 3, and 4 respectively. From the results, missing of KD or FKTM will lead to a significant loss of performance and robustness to stragglers, which validates the effectiveness of key components of MFL-AKD.

6 Conclusion

In this paper, we make the first attempt to analyze and combat stragglers in multimodal FL by integrating knowledge distillation, especially for vision-language scenarios. We propose a novel straggler-robust multimodal FL method named MFL-AKD. With the knowledge distilled from the vision-language pretrained model, the cross-modal semantic representations of the local model are greatly and rapidly enhanced, resulting in a remarkable improvement in convergence speed. Moreover, we also design the Fast Knowledge Transfer Mechanism to allow the balance of workload and enable early training of local models to handle the stragglers. Through experiments on video moment retrieval and text-image retrieval datasets, we have verified that our method can significantly alleviate the impact of stragglers while achieving remarkable retrieval accuracy.

Acknowledgments

This work was supported in part by the Alibaba Group through Alibaba Innovative Research Program, No.21169774; in part by the National Natural Science Foundation (NSF) of China, No.62206156, No.62276155, No.72004127, and No.62206157; in part by the NSF of Shandong Province, No.ZR2021MF040 and No.ZR2022QF047; in part by the Key R&D Program of Shandong Province, China (Major Scientific and Technological Innovation Projects), No.2022CXGC020107.

References

[Ahn *et al.*, 2020] Sungyong Ahn, Seunghyun Yoo, and Kijung Shin. Fedkd: Collaborative learning with weight-sharing and knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4598–4605, 2020.

[Bonawitz *et al.*, 2017] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Alec Segal, Karn Seth, Vincent Vanhoucke, et al. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[Cheng *et al.*, 2021] Sijie Cheng, Jingwen Wu, Yanghua Xiao, and Yang Liu. Fedgems: Federated learning of larger server models via selective knowledge fusion. *CoRR*, abs/2110.11027, 2021.

[Gao *et al.*, 2017] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017.

[Han *et al.*, 2020] Pengchao Han, Shiqiang Wang, and Kin K Leung. Adaptive gradient sparsification for efficient federated learning: An online learning approach. In *2020 IEEE 40th international conference on distributed computing systems (ICDCS)*, pages 300–310. IEEE, 2020.

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

[Hu *et al.*, 2023] Yupeng Hu, Kun Wang, Meng Liu, Haoyu Tang, and Liqiang Nie. Semantic collaborative learning for cross-modal moment localization. *ACM Transactions on Information Systems*, 42(2):1–26, 2023.

[Huang *et al.*, 2022] Jie Huang, Zheng Liu, Dong Yang, and Yang Li. Multimodal federated learning via contrastive representation ensemble. In *AAAI*, 2022.

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

[Komodakis and Pesquet, 2017] Nikos Komodakis and Jean-Christophe Pesquet. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5294–5303, 2017.

[Krishna *et al.*, 2017] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[Li *et al.*, 2020a] Mingsheng Li, Haoqiang Fan, Haoyi Xiong, Wei Zhu, Jianfei Hu, and Ling Li. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2020c] Xiang Li, Yifeng Wang, and Xiaoming M. Wu. Fedmd: Heterogenous federated learning via model distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2423–2433, 2020.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2019] Ronghang Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Elnaz Frank, Alexander Sergeev, and Kellen Yocum. Federated learning for vision-and-language grounding problems. In *EMNLP-IJCNLP 2019*, 2019.

[Liu *et al.*, 2021] Jianchun Liu, Hongli Xu, Lun Wang, Yang Xu, Chen Qian, Jinyang Huang, and He Huang. Adaptive asynchronous federated learning in resource-constrained edge computing. *IEEE Transactions on Mobile Computing*, 2021.

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

[Qu *et al.*, 2021] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Reisizadeh *et al.*, 2022] Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity. *IEEE Journal on Selected Areas in Information Theory*, 3(2):197–205, 2022.

[Romero *et al.*, 2015] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

[Smith *et al.*, 2017] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

[Tang *et al.*, 2021a] Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. Frame-wise cross-modal matching for video moment retrieval. *IEEE Transactions on Multimedia*, 24:1338–1349, 2021.

[Tang *et al.*, 2021b] Haoyu Tang, Jihua Zhu, Lin Wang, Qinghai Zheng, and Tianwei Zhang. Multi-level query interaction for temporal language grounding. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25479–25488, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021] Yuzhou Wang, Wenjie Jiang, Jingkuan Song, and Xiangyang Xue. Fedvmr: A new federated learning method for video moment retrieval. *IEEE Transactions on Multimedia*, 2021.

[Wu *et al.*, 2020] Wentai Wu, Ligang He, Weiwei Lin, Rui Mao, Carsten Maple, and Stephen Jarvis. Safa: A semi-asynchronous protocol for fast federated learning with low overhead. *IEEE Transactions on Computers*, 70(5):655–668, 2020.

[Wu *et al.*, 2022] Yuezhou Wu, Yan Kang, Jiahuan Luo, Yuanqin He, Lixin Fan, Rong Pan, and Qiang Yang. Fedcg: Leverage conditional gan for protecting privacy and maintaining competitive performance in federated learning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2334–2340. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Yang *et al.*, 2020] Qiang Yang, Yang Liu, Tianjian Chen, and Yuan Tong. *Federated Learning*, volume 14 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2020.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[Zhao *et al.*, 2021] Liang Zhao, Junlin Wang, Zhigang Chen, Qiang Wu, and Jie Chen. Multimodal federated learning on iot data. *IEEE Internet of Things Journal*, 2021.