FedDAT: An Approach for Foundation Model Finetuning in Multi-Modal Heterogeneous Federated Learning

Haokun Chen[1,2], Yao Zhang[1,4], Denis Krompass[2], Jindong Gu[3*], Volker Tresp[1,4]
[1] LMU Munich, Munich, Germany
[2] Siemens AG, Munich, Germany
[3] University of Oxford, Oxford, England
[4] Munich Center for Machine Learning (MCML), Munich, Germany
{haokun.chen, denis.krompass}@siemens.com, yzhang@dbs.ifi.lmu.de
jindong.gu@outlook.com, volker.tresp@lmu.de

Abstract

Recently, foundation models have exhibited remarkable advancements in multi-modal learning. These models, equipped with millions (or billions) of parameters, typically require a substantial amount of data for finetuning. However, collecting and centralizing training data from diverse sectors becomes challenging due to distinct privacy regulations. Federated Learning (FL) emerges as a promising solution, enabling multiple clients to collaboratively train neural networks without centralizing their local data. To alleviate client computation burdens and communication overheads, previous works have adapted Parameter-efficient Finetuning (PEFT) methods for FL. Hereby, only a small fraction of the model parameters are optimized and communicated during federated communications. Nevertheless, most previous works have focused on a single modality and neglected one common phenomenon, i.e., the presence of data heterogeneity across the clients. Therefore, in this work, we propose a finetuning framework tailored to heterogeneous multi-modal FL, called Federated Dual-Aadapter Teacher (FedDAT). Specifically, our approach leverages a Dual-Adapter Teacher (DAT) to address data heterogeneity by regularizing the client local updates and applying Mutual Knowledge Distillation (MKD) for an efficient knowledge transfer. FedDAT is the first approach that enables an efficient distributed finetuning of foundation models for a variety of heterogeneous Vision-Language tasks. To demonstrate its effectiveness, we conduct extensive experiments on four multi-modality FL benchmarks with different types of data heterogeneity, where FedDAT substantially outperforms the existing centralized PEFT methods adapted for FL.

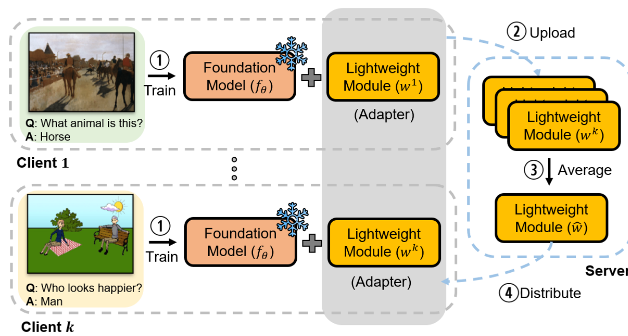


Figure 1: Schematic illustration of the training procedure for Visual Question Answering (VQA) in Federated Learning.

Introduction

Recent works have shown the power of foundation models with millions (billions) of parameters (Zhou et al. 2023; Du et al. 2022). These models, represented by Transfomers (Vaswani et al. 2017), achieve promising results when finetuned for real-world multi-modal tasks, including Visual Question Answering (VQA) (Antol et al. 2015), Visual Commonsense Reasoning (VCR) (Zellers et al. 2019), etc. To improve the generalization ability of the foundation models, a substantial amount of data from diverse sectors and application scenarios is typically required for extensive finetuning. However, it becomes challenging to aggregate all training data and perform centralized model finetuning. For instance, collecting data from different clinical centers across multiple countries becomes infeasible due to distinct privacy regulations, such as GDPR in the EU and PDPA in Singapore.

To address this problem, Federated Learning (FL) emerges as a promising solution, which allows a shared model to be collaboratively optimized using decentralized data sources. In the classical FL approaches, e.g., FedAvg (McMahan et al. 2017), the central server obtains the model by iteratively averaging the optimized model weights uploaded from the active clients. FL offers several advantages, including improved efficiency in client-server communication and enhanced data confidentiality, as it eliminates the need for direct access to the client's local dataset. FL provides promising solutions for various application areas, such as healthcare (Sheller et al. 2020) and industry (Liu et al. 2020), where data privacy is crucial.

Despite its promising prospects, traditional FL is unsuitable for finetuning the entire foundation model. The optimization and transmission of billions of parameters would impose significant client computation burdens and substantial communication overheads. To overcome this challenge,

*Corresponding author

parameter-efficient finetuning (PEFT) methods provide a possible solution, where only a small fraction of the model parameters is optimized and communicated during FL.

Existing works have predominantly explored a basic combination of centralized PEFT algorithms and FedAvg. For instance, some approaches focus on training and communicating only the tiny adaptation modules (adapter) (Houlsby et al. 2019; Su et al. 2022) or a small amount of trainable input tokens (Guo et al. 2022; Guo, Guo, and Wang 2023). However, these investigations are limited to single modality scenarios, where only visual or textual tasks are considered. Most importantly, none of these works address the problem of data heterogeneity, in which the data of different clients are not independent and identically distributed (*non-IID*). Data heterogeneity may lead to model drifts during the client local update, as well as an unstable and sub-optimal convergence of the aggregated server model (Li et al. 2020a; Mendieta et al. 2022). Therefore, in this paper, we propose Federated Dual-Adapter Teacher (`FedDAT`), as the first framework to address this challenging yet practical problem, PEFT of foundation models for multi-modal (Vision-Language) heterogeneous FL.

`FedDAT` incorporates a global adapter in the foundation model, which is optimized and transmitted during federated communications. `FedDAT` utilizes a Dual-Adapter Teacher (*DAT*) module, comprising two parallel adapters: one is a copy of the global adapter, kept frozen, while the other is locally optimized at each client. This configuration enables the local adapter to capture client-specific knowledge, which serves to regularize the global adapter and address data heterogeneity. Meanwhile, the frozen adapter preserves client-agnostic knowledge, thereby mitigating the catastrophic forgetting of the global adapter during knowledge transfer. To prevent overfitting of *DAT* to the limited client local dataset, we implement Mutual Knowledge Distillation (*MKD*) between *DAT* and the global adapter. This mechanism ensures efficient knowledge transfer while maintaining the generalization ability of both modules.

The proposed method `FedDAT` achieves state-of-the-art results on four multi-modality benchmarks that include a variety of Vision-Language (VL) tasks with data heterogeneity. Our contributions can be summarized as follows:

- We propose a novel method `FedDAT` for multi-modal heterogeneous FL, which is the first FL framework addressing distributed PEFT of foundation models for Vision-Language tasks.
- We conduct comprehensive experiments on four heterogeneous FL benchmarks with a variety of Vision-Language tasks. The results demonstrate that `FedDAT` achieves SOTA results, indicating better convergence rate and scalability compared to existing PEFT methods.

Related Work

Parameter-Efficient Finetuning (PEFT) for Federated Learning. PEFT has been well studied in centralized machine learning (Houlsby et al. 2019; Liu et al. 2022; Sung, Cho, and Bansal 2022), while its application on FL remains under-explored. Most of the prior work rudimentarily adapted PEFT for FL and focused on single-modal tasks:

(1) Image classification. (Chen et al. 2022; Sun et al. 2022) evaluate the existing PEFT baselines combined with FL, while (Guo et al. 2022; Guo, Guo, and Wang 2023; Li et al. 2023; Lu et al. 2023) finetune the CLIP model (Radford et al. 2021) via tuning and communicating only small amount of learnable (personalized) prompts. (Su et al. 2022) addresses the problem of heterogeneous client images by injecting lightweight adaptation modules (adapters) (Houlsby et al. 2019). (Yang et al. 2023) explores the possibility of finetuning generative foundation models (diffusion models) (Dhariwal and Nichol 2021) via FL.

(2) Language tasks. (Yu, Muñoz, and Jannesari 2023) requires public server dataset and optimize adapter for few-shot finetuning of BERT-like language models (Devlin et al. 2018). (Zhang et al. 2023) builds a distributed instruction tuning (Wei et al. 2021) datasets and finetunes the language model via Low-Rank Adaptation (LoRA) (Hu et al. 2021). (Zhuang, Chen, and Lyu 2023) systematically analyzes the challenges of finetuning large language models in FL.

(Yu et al. 2023) is the first to analyze the situation of having multi-modal client datasets and conducts contrastive representation learning. However, the visual data and the language data are processed by separate networks, i.e., no Vision-Language Foundation Model is involved. In this work, we focus on the under-explored PEFT for large-scale vision-language models in FL and address the problem of client local datasets with heterogeneity in both vision and/or language modality.

Vision-Language Foundation Model. Vision-Language foundation models have significantly advanced the Vision-Language tasks (Antol et al. 2015; Zellers et al. 2019; Suhr et al. 2019; Xie et al. 2019a). Based on the perspective of intra-modality data handling, there are two types of mainstream Vision-Language Foundation model structures: (1) Single-stream Vision-Language Foundation models (Li et al. 2019; Chen et al. 2020; Li et al. 2020b; Su et al. 2020; Kim, Son, and Kim 2021a; Singh et al. 2022), which directly fuse the initial language/visual representation by using the joint cross-modal encoder at the initial state, and (2) Dual-stream Vision-Language foundation models (Lu et al. 2019; Tan and Bansal 2019; Li et al. 2021b; Huo et al. 2021), which separately apply the intra-modality processing to two modalities along with a shared cross-modal encoder. To showcase the applicability of our proposed `FedDAT` to a wide range of Vision-Language foundation models, we carefully select ViLT (Kim, Son, and Kim 2021a) as a representative single-stream Vision-Language foundation model, and ALBEF (Li et al. 2021b) as a representative dual-stream Vision-Language foundation model. By employing these diverse models, we effectively demonstrate the versatility and robustness of `FedDAT` in Vision-Language learning.

Methodology

Problem Statement

In this work, we address a heterogeneous FL problem setting with K clients: Each client k owns its private multi-modal dataset D^k, containing data from visual modality (images) and textual modality (texts). Specifically, we focus on the vision-language tasks and take Visual Question Answering (VQA) as an example. Hereby, the local dataset D^k can be further decomposed into N_k image-question-answer triplets $\{(v_i^k, q_i^k, a_i^k)|i \in \{1, ..., N_k\}\}$. We assume that the marginal distribution of v_i^k and/or q_i^k, a_i^k varies across the clients, i.e., there exists data heterogeneity in the visual space and/or in the textual space. We define the answer pool $A^k = \{a_1^k, ..., a_{C^k}^k\}$ with C^k ground-truth answers for client k and define our task as a C^k-way classification problem following (Antol et al. 2015). Note that the size of the answer pool could be different for different clients. The objective of FL is to collaboratively finetune one *global* foundation model f_θ in a parameter-efficient manner (PEFT) within a pre-defined communication budget, which produces promising results on all client's local data.

PEFT Method: Adapter

In this section, we introduce a traditional parameter-efficient finetuning (PEFT) method, i.e., Adapter (Houlsby et al. 2019), adjusted for FL applications. Here, we adopt the foundation models with common Transformer architecture (Vaswani et al. 2017) consisting of multiple repeated Transformer blocks. Specifically, each block contains a self-attention sub-layer, a fully connected feed-forward network (FFN), and residual connections around the sub-layers followed by layer normalization.

Adapter is a bottleneck network consisting of a down-sample linear layer $W_{down} \in \mathbb{R}^{d\times r}$ and an up-sampling linear layer $W_{up} \in \mathbb{R}^{r\times d}$, where r denotes the down-sampled dimension ($r < d$). A nonlinear activation function $\phi(\cdot)$, such as ReLU, is inserted in between. The adapter is injected after the FFN of each Transformer block and its computation can be formulated as

$$h' = h + \phi(hW_{down})W_{up}, \tag{1}$$

where h is the normalized output of FFN.

Recap: Federated Averaging

In this section, we formally describe the combination of the conventional federated learning algorithm, FedAvg (McMahan et al. 2017), and the centralized PEFT algorithm, i.e., Adapter. Before the client-server communication starts, we deploy the *same* pre-trained foundation model f_θ at different clients. Afterwards, the server randomly initializes the parameter w of the learnable lightweight module, which are the weight matrices of the linear layers W_{down} and W_{up} in the adapters. w is then distributed to all clients for communication and local optimization. We illustrate the procedure of one communication round in the following.

As shown in Figure 1, each active client k first execute local training to optimize the light-wegiht module w^k combined with the *frozen* foundation model f_θ (①) in parallel, where the following loss L_k is minimized:

$$L_k(w^k) = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathcal{L}(y_i, f_{\theta\cup w^k}(x_i)), \tag{2}$$

where y_i is the ground-truth label of input data x_i, and $\mathcal{L}$ is the loss function, e.g., Cross-Entropy for classification tasks. After the local updates, the central server aggregates $\{w^k|1 \le k \le K\}$, uploaded (②) by all active clients, and executes a parameter aggregation (③):

$$\hat{w} \leftarrow \frac{1}{\sum_{k=1}^{K} N_k}\sum_{k=1}^{K} N_k \cdot w^k. \tag{3}$$

Finally, the aggregated weight $\hat{w}$ will be distributed (④) to the active clients for optimization in the next communication round. Note that after exhausting all communication budgets, the global model $f_{\theta\cup w}$ is deployed for the testing.

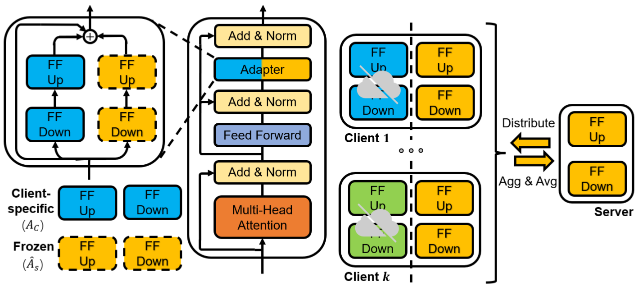
Motivational Case Study

To motivate the architecture design of `FedDAT`, we present an empirical analysis to address the following research question: *Which type of knowledge is more crucial for optimizing a promising ML model in heterogeneous FL, client-specific or client-agnostic?* Therefore, we follow the experiment design proposed in (Tan et al. 2022). Specifically, we take the down-sampled version of DomainNet (Peng et al. 2019), which is an image classification benchmark and contains data from 6 different styles: Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R), and Sketch (S). By assigning data from one style to each client, we simulate data heterogeneity in the feature space across different clients. We finetune the foundation model, i.e., ViT (Dosovitskiy et al. 2020), with different PEFT methods via FL.

In Table 1, we provide the results of finetuning the classification head (clf) and finetuning with $Adapter$. We also display the performance of client local finetuning (L), i.e., no federated communication involved. We conclude three observations from the results: (1) $Adapter$ is an effective PEFT method in both federated setting and independent finetuning setting compared with clf, providing an average performance increase of 3.22% and 6.15%, respectively. (2) Collaborative training via FL, i.e., finetuning a *client-agnostic* foundation model, generally outperforms local independent finetuning. This can be observed by comparing the average accuracy of models with and without "L". (3) *Client-specific* classification head and adapters show benefits on certain clients (marked with underlines), i.e., clients with Painting (P) and Sketch (S) data and optimized independently. We assume this is due to the large distribution shift in the feature space across different clients' local data, given their different image appearances. This phenomenon

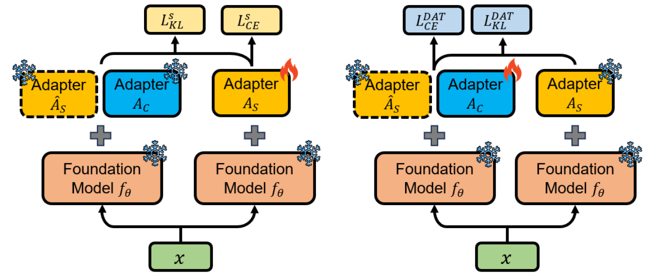
Method	DomainNet						
	C	I	P	Q	R	S	avg
clf-L	72.43	36.13	<u>86.35</u>	55.70	74.07	<u>74.70</u>	66.56
$Adapter$-L	76.05	36.93	<u>88.03</u>	72.40	66.53	<u>78.74</u>	69.78
clf	80.80	44.61	<u>83.47</u>	60.10	84.21	<u>71.69</u>	70.81
$Adapter$	88.59	50.95	<u>87.12</u>	76.00	84.99	<u>74.08</u>	**76.96**

Table 1: Evaluation results of ViT finetuned for DomainNet with/without FL. "L" indicates independent client training, i.e., no federated communication involved.



(a) *DAT* Architecture Design (b) Communication

Figure 2: Schematic illustration of the Dual-Adapter Teacher (*DAT*) with local A_c and frozen $\hat{A}_s$. Only the shared adapter A_s is transmitted during federated communication.



(a) Optimization of A_s (b) Optimization of *DAT*

Figure 3: Schematic illustration of the Mutual Knowledge Distillation (*MKD*) between *DAT* and A_s.

answers the previous research question: Both *client-specific* and *client-agnostic* knowledge are crucial and should not be forgotten during federated communication. These observations motivate the proposed method and serve as evidence for its promising applicability and effectiveness.

Proposed Method

In this section, we introduce the proposed method Federated Dual-Adapter Teacher (`FedDAT`). As shown in Algorithm 1, the training process of `FedDAT` can be divided into two functions, which will be introduced in the following:

At the beginning of the training, the server initializes a shared adapter A_s. In each communication round, all active clients receive A_s and conduct $Client\ Update$ in parallel. Subsequently, the server aggregates and averages the optimized parameters $\{A_s^k|1 \leq k \leq K\}$ uploaded from all clients, which will be used as the initialization of A_s for the next communication round.

The client local update comprises 2 main components, which will be introduced in the following:

(1) Dual-Adapter Teacher (*DAT*). Before the first communication round, each client locally initializes the local adapter A_c as well as the foundation model f_θ with the same pre-trained weights θ. Subsequently, each client receives the parameters of A_s from the server, which is then copied as $\hat{A}_s$ and kept frozen during the client local update. We combine $\hat{A}_s$ and A_c as the Dual-Adapter Teacher (*DAT*) and provide its schematic illustration in Figure 2a.

In *DAT*, we constrain the parameters of A_c strictly local for each client. By personalizing A_c, we force it to focus solely on client-specific knowledge, which is crucial for client data heterogeneity. Meanwhile, the frozen $\hat{A}_s$ is utilized to retain the client-agnostic knowledge captured by the shared adapter A_s. Similar to traditional adapters (Equation 1), given the normalized output of FFN h in a Transformer layer, *DAT* performs the following transformation:

$$h' \leftarrow h + \frac{1}{2}\phi(h \cdot \hat{W}_s^{down}) \cdot \hat{W}_s^{up} + \frac{1}{2}\phi(h \cdot W_c^{down}) \cdot W_c^{up}, \quad (4)$$

where $\hat{W}_s$ and W_c are the weight matrices for $\hat{A}_s$ and A_c, respectively. Afterwards, T local update steps will be executed, in which the shared adapter A_s and the *DAT* module is optimized.

By utilizing *DAT* as a guidance for the local optimization of A_s at each client, our goal is to distill client-specific knowledge into A_s and mitigate the forgetting of A_s on its client-agnostic knowledge. Hereby, we apply Mutual Knowledge Distillation (*MKD*) for an efficient knowledge transfer, which will be introduced in the following.

(2) Mutual Knowledge Distillation (*MKD*). A schematic illustration of *MKD* is provided in Figure 3. *MKD* executes bi-directional knowledge distillation between A_s and *DAT* via L^s_{KL} and $L^{\text{DAT}}_{\text{KL}}$, respectively:

$$L^s_{\text{KL}} = \mathcal{KL}(z_s(x)||z_{\text{DAT}}(x)), \quad L^{\text{DAT}}_{\text{KL}} = \mathcal{KL}(z_{\text{DAT}}(x)||z_s(x)), \quad (5)$$

where $\mathcal{KL}$ denotes the Kullback-Leibler divergence, z_s and z_{DAT} are the predicted logits of the foundation model injected with A_s and *DAT*, respectively. Hereby, this setup allows the shared adapter A_s to capture both client-specific knowledge and client-agnostic stored in *DAT* (L^s_{KL}). Additionally, we apply A_s as guidance for the optimization *DAT* ($L^{\text{DAT}}_{\text{KL}}$) to prevent possible overfitting, considering the scarce local data of each client (McMahan et al. 2017).

MKD is utilized together with the guidance from ground-truth labels of the training data, i.e.,

$$\begin{aligned} L^s_{\text{CE}} &= \sum_{c=1}^{C} \mathcal{I}(x,c) \cdot log(\sigma(z_s(x))^{(c)}), \\ L^{\text{DAT}}_{\text{CE}} &= \sum_{c=1}^{C} \mathcal{I}(x,c) \cdot log(\sigma(z_{\text{DAT}}(x))^{(c)}), \end{aligned} \quad (6)$$

where, $\mathcal{I}(x,c)$ is a binary indicator (0 or 1) if c is the ground-truth label for x, σ is the softmax function. Hereby, we aim at training the foundation model, injected with either A_s or *DAT*, to correctly classify the training sample x. Finally, combining *MKD* and L_{CE} produces the optimization objective for A_s and *DAT*:

$$\begin{aligned} L^s &= L^s_{\text{CE}} + \alpha L^s_{\text{KL}}, \\ L^{\text{DAT}} &= L^{\text{DAT}}_{\text{CE}} + \beta L^{\text{DAT}}_{\text{KL}}, \end{aligned} \quad (7)$$

where, α and β are the weighting coefficient. While both *DAT* and A_s are randomly initialized, they become more informative as the training progresses. To reflect this observation, we apply an exponential ramp-up schedule for α and β. Despite the sophisticated design of our method, `FedDAT` indicates the same inference cost and communication overhead as the PEFT method $Adapter$, where only A_s is transmitted and applied at deployment.

Experiments and Analyses

We conduct extensive empirical analyses to investigate the proposed method. Firstly, we compare `FedDAT` with other centralized PEFT methods on four heterogeneous FL benchmarks containing different Vision-Language tasks. Afterwards, we demonstrate the effectiveness of `FedDAT` components via ablation study. Finally, we analyze the promising convergence rate and scalability of `FedDAT`.

Benchmark Experiments

Datasets Description. We conduct experiments on different Vision-Language (VL) benchmarks with different types of data heterogeneity, including visual, textual, and task heterogeneity. We introduce these benchmarks in the following.

- **Domain**. We adopt 5 common VQA datasets from different domains, i.e., VizWiz (Gurari et al. 2018), COCO QA (Ren, Kiros, and Zemel 2015), Art (Garcia et al. 2020), GQA (Hudson and Manning 2019) and Abstract (Antol et al. 2015). We assign one of the datasets to each client, leading to heterogeneity in both vision and language modality. Example VQA triplets from the benchmark are provided in Figure 4.
- **Function** & **Scene**. We adopt and split the CLOVE benchmark (Lei et al. 2023) into *Scene* and *Function* benchmark, which contains VQA triplets collected from 6 different visual environments and 5 different functions, respectively. Triplets from one scene (function) are allocated to one client, resulting in visual (textual) heterogeneity in the *Scene* (*Function*) benchmark.
- **Task**. We adopt and modify the CLiMB benchmark (Srinivasan et al. 2022), which contains 4 VL tasks, namely VQA (Antol et al. 2015), Natural Language for Visual Reasoning (*NLVR*) (Suhr et al. 2018), Visual Entailment (*VE*) (Xie et al. 2019b), and Visual Commonsense Reasoning (*VCR*) (Zellers et al. 2019). Each client owns data from one of the datasets, introducing task heterogeneity across different clients.

We downsample the original dataset to simulate client local data scarcity described in prior arts (McMahan et al. 2017) and provide more details in the Appendix.

Implementation Details. For the task-heterogeneous benchmark (*Task*), we adopt the Transformer encoder-only backbones following (Srinivasan et al. 2022), i.e., ViLT (Kim, Son, and Kim 2021b) and VAuLT (Chochlakis et al. 2022). For the rest three benchmarks, we add another encoder-decoder backbone, i.e., ALBEF (Li et al. 2021a). We compare `FedDAT` with various centralized PEFT methods adapted for FL, including $LoRA$ (Hu et al. 2021), $prompt$-tuning (Guo et al. 2022), and $bias$-tuning (Cai et al.

Algorithm 1: Training procedure of FedDAT

ServerUpdate
```
Randomly initialize A_s
for round r = 1 to R do
    for client k = 1 to K do {in parallel}
        A_s^k ← ClientUpdate(A_s, k, r)
    end for
    A_s ← (1/K) Σ_{k=1}^{K} A_s^k
end for
```
ClientUpdate(A_s, k, r)
```
if r = 1 then
    Randomly initialize A_c
end if
Â_s ← A_s
for local step t = 1 to T do
    Sample {X, y} from D_k
    Optimize A_s via minimizing L^s
    Optimize DAT via minimizing L^DAT
end for
return A_s
```

2020). We also provide results of independent client optimization (marked by "L") of the classification head clf and $Adapter$. Moreover, we provide the results of fully finetuning the models ($full$) as an *oracle* method (marked by $*$), given the infeasibility of transmitting the entire foundation model in FL.

To handle the different answer pools in different clients, we incorporate client-specific classification heads for ViLT and VAuLT, and apply client-specific answer lists for ALBEF. To make a fair comparison between different centralized PEFT algorithms and `FedDAT`, we apply the same hyperparameters search for all methods in different benchmarks. All experiments are repeated with 3 random seeds. The hyperparameters are detailed in the Appendix.

Results and Analyses. In Table 2, we provide the results of `FedDAT` and the other FL-adapted PEFT methods on our *Domain* benchmark. We observe `FedDAT` outperforms all the baselines with all the architectures, achieving an average performance improvement of up to 4.55% compared with the most promising baseline $Adapter$. This indicates the easy adaptability of `FedDAT` for both encoder-based and encoder-decoder-based VL models. Moreover, `FedDAT` depicts the same communication overhead as a single $Adapter$, which adds and optimizes only less than

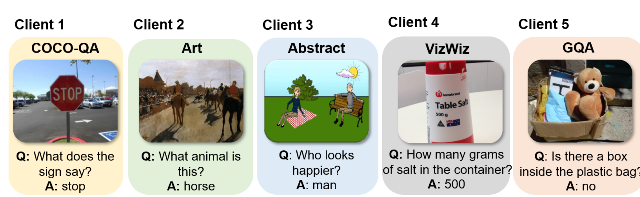


Figure 4: Example VQA triplets of different datasets in *Domain* benchmark with heterogeneity in both Vision and Language modality.

Backbone	Method	Comm. Overhead	VizWiz	COCO	Art	GQA	Abstract	Average
ViLT	clf-L	–	63.13±1.07	36.15±2.92	63.22±0.99	34.90±3.16	52.81±2.67	50.04±1.81
	$LoRA$	$0.60M(0.48\%)$	60.47±1.25	43.28±1.37	62.98±0.75	36.57±2.01	52.04±1.62	51.07±1.41
	$prompt$	$0.60M(0.48\%)$	60.13±1.05	52.13±0.87	63.02±1.58	39.09±0.37	52.88±3.07	53.45±2.04
	$bias$	$0.10M(0.08\%)$	61.83±2.41	49.41±2.36	69.38±1.69	40.43±0.66	60.36±1.92	56.28±1.97
	$Adapter$-L	–	61.72±1.42	46.27±4.58	67.69±0.42	43.62±0.93	54.02±2.16	54.67±2.54
	$Adapter$	$0.89M(0.75\%)$	61.39±1.11	52.39±6.20	68.72±3.20	43.72±0.65	59.43±2.94	57.13±4.08
	FedDAT	$0.89M(0.75\%)$	60.99±2.81	63.81±2.90	71.36±3.34	48.65±2.93	60.75±2.67	**61.11**±2.98
	$full$-L*	–	55.52±1.42	72.97±1.53	73.16±0.28	44.41±3.98	58.78±0.25	60.97±1.45
	$full$*	$87.40M(100\%)$	56.12±2.55	73.87±0.83	76.24±1.82	50.28±1.59	61.26±0.78	63.55±1.35
VAuLT	clf-L	–	61.83±1.85	32.42±0.04	64.52±1.55	35.08±5.57	48.48±0.77	48.46±1.15
	$LoRA$	$0.60M(0.29\%)$	62.17±1.32	40.56±0.86	63.08±1.13	33.47±3.08	47.34±1.04	49.32±1.16
	$prompt$	$0.60M(0.29\%)$	62.93±0.87	46.52±1.45	64.26±1.03	35.33±2.12	48.91±0.68	51.59±1.63
	$bias$	$0.21M(0.10\%)$	61.12±2.84	43.81±0.35	67.00±1.41	33.30±4.81	51.22±2.07	51.29±1.08
	$Adapter$-L	–	62.33±1.42	47.72±2.83	67.50±2.11	33.75±2.79	54.09±0.93	53.07±1.34
	$Adapter$	$1.79M(0.77\%)$	52.53±3.65	53.63±0.28	66.80±0.53	35.65±1.84	50.03±1.77	51.73±0.46
	FedDAT	$1.79M(0.77\%)$	62.19±1.01	54.83±2.04	67.86±1.93	40.06±3.08	54.48±0.49	**55.88**±1.79
	$full$-L*	–	57.41±2.13	55.68±1.24	70.27±2.11	41.31±1.46	52.66±0.57	55.47±1.85
	$full$*	$227.77M(100\%)$	45.79±2.12	64.64±3.05	67.89±1.82	41.93±3.85	49.58±0.66	53.97±2.09
ALBEF	$LoRA$	$1.52M(0.53\%)$	60.49±1.32	28.32±0.65	57.04±3.69	28.71±0.42	58.06±2.42	46.52±1.75
	$prompt$	$0.92M(0.32\%)$	63.13±0.65	32.50±1.20	63.45±0.42	32.08±1.07	59.45±1.78	50.12±0.95
	$bias$	$0.93M(0.32\%)$	63.23±0.14	31.23±0.28	61.23±1.12	35.93±1.73	57.88±0.28	49.90±0.87
	$Adapter$-L	–	61.72±1.12	56.32±1.50	65.21±0.35	40.96±2.27	59.51±1.58	56.74±1.38
	$Adapter$	$2.86M(0.98\%)$	59.52±2.44	69.35±2.78	68.32±0.89	41.02±3.12	60.83±2.66	59.81±1.87
	FedDAT	$2.86M(0.98\%)$	61.52±1.51	76.36±0.63	71.04±0.50	49.22±1.60	63.65±1.19	**64.36**±1.39
	$full$-L*	–	61.22±0.14	77.80±1.39	74.45±0.7	50.09±1.06	63.58±2.79	65.43±1.37
	$full$*	$290.34M(100\%)$	51.91±1.42	78.38±1.11	75.65±0.14	55.91±0.54	70.47±0.83	66.46±0.96

Table 2: Evaluation results of different finetuning methods on our FL benchmark with distribution shift in both Vision and Language space. "L" indicates client local finetuning where no communication is involved. We report the mean±std accuracy of each client from 3 runs with different seeds.

1% of the total parameters in the foundation model. This further illustrates its applicability to the FL system with constrained communication bandwidths. Besides, `FedDAT` narrows the performance gap between the PEFT methods and fully-finetuning methods. Interestingly, our approach outperforms the oracle methods $full$-L when applied on ViLT and VAuLT, which demonstrates the effectiveness of introducing client-specific knowledge into the client local optimization. We also note that applying $Adapter$-L for VAuLT, i.e., optimizing adapters for each client independently, achieves better results than $Adapter$, which provides additional evidence for our observation in Section .

Afterwards, we provide the comparison of clients' average accuracy between `FedDAT` and different PEFT methods on the other benchmarks. As shown in Table 3, `FedDAT` provides promising improvements of up to 6.02%, 7.94% and 1.09% on *Function, Scene*, and *Task* benchmark, respectively. More details of the client specific performance are provided in the Appendix.

Ablation Study

To illustrate the importance of different components used in `FedDAT`, we conduct an ablation study for ViLT on three benchmarks. The results are shown in Table 4. We first investigate the optimization process, where we notice that optimizing without *DAT*, i.e., applying solely the local adapter A_c or the frozen adapter $\hat{A}_s$ as the teacher, leads to only minimal performance increase, which indicates the effectiveness of our Dual-Adapter Teacher design. Besides, distilling only the knowledge from *DAT* to the shared adapter A_s, i.e., omitting the bi-directional *MKD*, brings visible performance gain. Combining both strategies achieves the best results, which further demonstrates their complementarity. Additionally, we validate other inference choices. Specifically, we evaluate the final *DAT* module (combination of A_c and $\hat{A}_s$) and the local adapter A_c at each client. We again note that we are addressing the problem of finetuning a global foundation model via FL, where no further personalization is required. Considering the inference efficiency and the problem setting, we adopt the shared adapter A_s for inference, which also achieves the most promising results.

Convergence Analysis

In Figure 5, we display the convergence analysis of `FedDAT` compared with the most promising PEFT method $Adapter$ on *Domain* benchmark. Hereby, we report the accuracy of the clients on their corresponding local testing set after each communication round. As shown in the figure, even though `FedDAT` utilizes a more sophisticated optimization schema, i.e., a combination of *DAT* and *MKD*, the learning curves of `FedDAT` still exhibit faster convergence rates than single $Adapter$. It is also worth noticing that `FedDAT` already achieves distinct performance gain after 5 communication rounds, i.e., 25% of the total communication budgets.

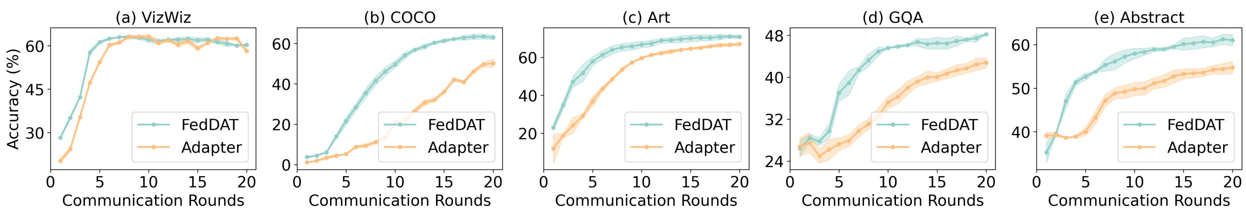


Figure 5: Convergence analysis of ViLT model on different clients in *Domain* benchmark.

Backbone	Method	Function	Scene	Task
ViLT	*clf*-L	31.58±1.97	24.52±0.95	49.46±0.39
	LoRA	32.04±1.12	28.47±1.03	47.82±1.42
	prompt	40.53±1.56	30.53±1.30	49.55±1.14
	bias	43.81±1.39	33.65±1.87	50.71±1.26
	Adapter-L	39.68±2.19	31.91±2.05	49.59±1.74
	Adapter	48.37±1.56	31.07±1.08	51.44±1.34
	FedDAT	**54.39**±2.36	**39.35**±1.25	**52.37**±0.52
	full-L*	56.81±2.97	38.00±1.48	50.64±1.42
	*full**	59.62±2.56	40.62±3.76	53.17±0.69
VAuLT	*clf*-L	27.72±3.05	21.22±2.08	39.63±1.07
	LoRA	29.87±1.86	23.08±1.09	38.35±1.47
	prompt	36.32±2.07	25.63±1.54	38.75±1.34
	bias	36.11±3.05	24.89±2.17	39.46±0.99
	Adapter-L	37.22±2.38	28.57±1.98	40.42±1.21
	Adapter	41.50±3.24	29.39±2.65	40.19±0.89
	FedDAT	**44.54**±2.08	**34.31**±2.87	**41.28**±0.57
	full-L*	49.13±2.68	35.11±1.99	41.66±1.32
	*full**	46.38±1.57	36.72±2.57	42.44±0.71

Table 3: Evaluation results of different methods on *Function*, *Scene*, and *Task* benchmark. "L" indicates independent client finetuning. We report the mean±std accuracy of 3 trials.

Scalability Analysis of `FedDAT`

To show the effectiveness of `FedDAT` under various application scenarios, we further conduct experiments with different numbers of clients. More specifically, we split the data of each function in the original CLOVE dataset (Lei et al. 2023) into 5 subsets, where each subset has an equal number of training data and is assigned to one client, following the client data scarcity described in (McMahan et al. 2017). We conduct experiments where 1, 2, 3, 4, and 5 clients (subsets) from each function are selected, which gives in total 5, 10, 15, 20, and 25 clients joining the federated communication for the *Function* benchmark, respectively. We apply also the same split strategy for the 6 different visual environments for the *Scene* benchmark and conduct the same experiment. More details regarding the experimental setups are provided in Appendix.

We observe that `FedDAT` consistently outperforms *Adapter* across all setups with small or large quantities of training data. Notably, a performance gap of up to 10% for ALBEF and 6% for ViLT is evident. These results indicate the scalability of `FedDAT` in handling complex FL applications involving a larger number of clients and increased communication budgets.

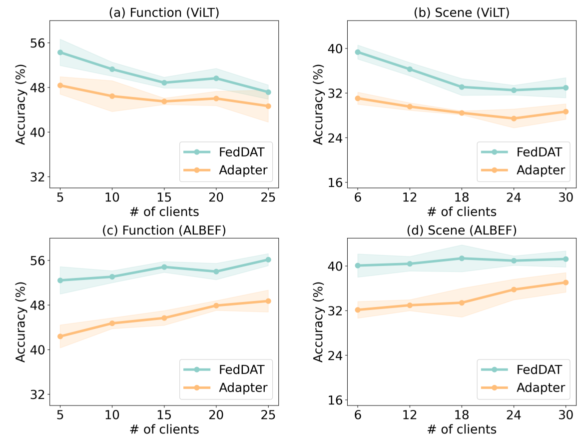


Figure 6: Scalability analysis of `FedDAT` with different number of clients on *Funciton* and *Scene* benchmarks.

Conclusion

In this work, we propose the first FL framework to address the parameter-efficient finetuning (PEFT) of the foundation model in heterogeneous FL, where various Vision-Language tasks are investigated. The proposed method, named `FedDAT`, optimizes a shared adapter utilizing the Dual-Adapter Teacher (*DAT*) and Mutual Knowledge Distillation (*MKD*). Compared with existing centralized PEFT methods, `FedDAT` achieves promising results on the four FL benchmarks with various Vision-Language tasks, demonstrating its effectiveness. Additional experiments indicate its applicability to complex FL setups involving larger distributed systems and training budgets.

Stage	Method	Domain	Function	Scene
-	*Adapter*	57.13±4.08	48.37±1.56	31.07±1.08
Optimization	w/o $\hat{A}_s$	58.24±0.98	50.62±1.45	33.04±0.65
	w/o A_c	57.87±1.24	50.93±0.85	32.45±0.27
	w/o *MKD*	58.41±1.57	52.82±2.98	36.98±1.07
	FedDAT	**61.11**±2.98	**54.39**±2.36	**39.35**±1.25
Inference	$A_c + A_s$	58.45±1.57	50.42±1.87	35.61±2.41
	A_c	55.87±3.35	46.14±2.60	32.84±0.78
	A_s (**FedDAT**)	**61.11**±2.98	**54.39**±2.36	**39.35**±1.25

Table 4: Ablation study for different components in optimization and inference stage of `FedDAT` on three benchmark datasets.

References

[3] Shen, Y.; and et al. 2022. Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Cai, H.; Gan, C.; Zhu, L.; and Han, S. 2020. Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33: 11285–11297.

Chen, J.; Xu, W.; Guo, S.; Wang, J.; Zhang, J.; and Wang, H. 2022. FedTune: A Deep Dive into Efficient Federated Fine-Tuning with Pre-trained Transformers. *arXiv preprint arXiv:2211.08025*.

Chen, S.; Gu, J.; Han, Z.; Ma, Y.; Torr, P.; and Tresp, V. 2023. Benchmarking Robustness of Adaptation Methods on Pre-trained Vision-Language Models. *arXiv preprint arXiv:2306.02080*.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, 104–120. Springer.

Chochlakis, G.; Srinivasan, T.; Thomason, J.; and Narayanan, S. 2022. Vault: Augmenting the vision-and-language transformer with the propagation of deep language representations. *arXiv preprint arXiv:2208.09021*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Liu, Z.; Li, J.; and Zhao, W. X. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Garcia, N.; Ye, C.; Liu, Z.; Hu, Q.; Otani, M.; Chu, C.; Nakashima, Y.; and Mitamura, T. 2020. A Dataset and Baselines for Visual Question Answering on Art. In *Proceedings of the European Conference in Computer Vision Workshops*.

Guo, T.; Guo, S.; and Wang, J. 2023. pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning. In *Proceedings of the ACM Web Conference 2023*, 1364–1374.

Guo, T.; Guo, S.; Wang, J.; and Xu, W. 2022. PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models–Federated Learning in Age of Foundation Model. *arXiv preprint arXiv:2208.11625*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *CVPR*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Huo, Y.; Zhang, M.; Liu, G.; Lu, H.; Gao, Y.; Yang, G.; Wen, J.; Zhang, H.; Xu, B.; Zheng, W.; et al. 2021. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Kim, W.; Son, B.; and Kim, I. 2021a. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 5583–5594. PMLR.

Kim, W.; Son, B.; and Kim, I. 2021b. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Le, H. Q.; Nguyen, M. N.; Thwal, C. M.; Qiao, Y.; Zhang, C.; and Hong, C. S. 2023. FedMEKT: Distillation-based Embedding Knowledge Transfer for Multimodal Federated Learning. *arXiv preprint arXiv:2307.13214*.

Lei, S. W.; Gao, D.; Wu, J. Z.; Wang, Y.; Liu, W.; Zhang, M.; and Shou, M. Z. 2023. Symbolic replay: Scene graph as prompt for continual learning on vqa task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1250–1259.

Li, G.; Wu, W.; Sun, Y.; Shen, L.; Wu, B.; and Tao, D. 2023. Visual Prompt Based Personalized Federated Learning. *arXiv preprint arXiv:2303.08678*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021b. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 9694–9705.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.; and Chang, K. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR*, abs/1908.03557.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020b. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, 121–137. Springer.

Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; and Raffel, C. A. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35: 1950–1965.

Liu, Y.; Garg, S.; Nie, J.; Zhang, Y.; Xiong, Z.; Kang, J.; and Hossain, M. S. 2020. Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach. *IEEE Internet of Things Journal*, 8(8): 6348–6358.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13–23.

Lu, W.; Hu, X.; Wang, J.; and Xie, X. 2023. FedCLIP: Fast Generalization and Personalization for CLIP in Federated Learning. *arXiv preprint arXiv:2302.13485*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mendieta, M.; Yang, T.; Wang, P.; Lee, M.; Ding, Z.; and Chen, C. 2022. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8397–8406.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1406–1415.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Sheller, M. J.; Edwards, B.; Reina, G. A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R. R.; et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1): 12598.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 15617–15629. IEEE.

Srinivasan, T.; Chang, T.-Y.; Pinto Alva, L.; Chochlakis, G.; Rostami, M.; and Thomason, J. 2022. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 35: 29440–29453.

Su, S.; Yang, M.; Li, B.; and Xue, X. 2022. Cross-domain Federated Adaptive Prompt Tuning for CLIP. *arXiv preprint arXiv:2211.07864*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 6418–6428. Association for Computational Linguistics.

Sun, G.; Mendieta, M.; Yang, T.; and Chen, C. 2022. Exploring Parameter-Efficient Fine-tuning for Improving Communication Efficiency in Federated Learning. *arXiv preprint arXiv:2210.01708*.

Sung, Y.-L.; Cho, J.; and Bansal, M. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5227–5237.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 5099–5110. Association for Computational Linguistics.

Tan, Y.; Long, G.; Ma, J.; Liu, L.; Zhou, T.; and Jiang, J. 2022. Federated learning from pre-trained models: A contrastive learning approach. *Advances in Neural Information Processing Systems*, 35: 19332–19344.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019a. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. *CoRR*, abs/1901.06706.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019b. Visual Entailment: A Novel Task for Fine-grained Image Understanding. *arXiv preprint arXiv:1901.06706*.

Yang, M.; Su, S.; Li, B.; and Xue, X. 2023. Exploring One-shot Semi-supervised Federated Learning with A Pre-trained Diffusion Model. *arXiv preprint arXiv:2305.04063*.

Yang, X.; Xiong, B.; Huang, Y.; and Xu, C. 2022. Cross-Modal Federated Human Activity Recognition via Modality-Agnostic and Modality-Specific Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3063–3071.

Yu, Q.; Liu, Y.; Wang, Y.; Xu, K.; and Liu, J. 2023. Multimodal Federated Learning via Contrastive Representation Ensemble. *arXiv preprint arXiv:2302.08888*.

Yu, S.; Muñoz, J. P.; and Jannesari, A. 2023. Federated Foundation Models: Privacy-Preserving and Collaborative Learning for Large Models. *arXiv preprint arXiv:2305.11414*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J.; Vahidian, S.; Kuo, M.; Li, C.; Zhang, R.; Wang, G.; and Chen, Y. 2023. Towards Building the Federated GPT: Federated Instruction Tuning. *arXiv preprint arXiv:2305.05644*.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Zhuang, W.; Chen, C.; and Lyu, L. 2023. When Foundation Model Meets Federated Learning: Motivations, Challenges, and Future Directions. *arXiv preprint arXiv:2306.15546*.